ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS

DEPARTAMENTO DE POSTGRADOS

TEMA:

“ANÁLISIS MULTIVARIANTE: SEGMENTACIÓN DE CLIENTES Y ESTRATEGIAS DE VENTAS DE UNA EMPRESA DE CONSUMO MASIVO DE GUAYAQUIL”

PROYECTO DE TITULACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MAGÍSTER EN ESTADISTICA APLICADA

AUTOR:

KEVIN JAVIER QUELAL ROBALINO

GUAYAQUIL - ECUADOR

2024

DEDICATORIA

Dedico este trabajo a Dios, por haberme guiado y bendecido durante este viaje que llamamos vida.

A mi padre por su amor incondicional y apoyo constante, que han sido fundamentales en mi vida.

A mi esposa ha sido mi compañera incansable, su fuerza, amor y comprensión han enriquecido cada aspecto de mi existencia.

A mi hija que es la luz de mis ojos.

A todos mis amigos y mentores que han dejado una huella imborrable, su compañerismo y apoyo han sido invaluables en los momentos más difíciles.

KEVIN QUELAL R.

AGRADECIMIENTOS

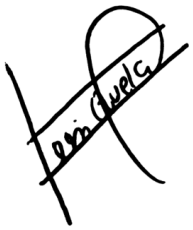
A mi mi esposa Nicole y mi hija Nyah por ser el motor para seguir adelante.
A mi tutora Andrea García por su soporte y enseñanzas para poder culminar este proyecto

KEVIN QUELAL R.

DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Titulación, me corresponde exclusivamente y ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente, este trabajo es de mi total autoría. El patrimonio intelectual del mismo, corresponde exclusivamente a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

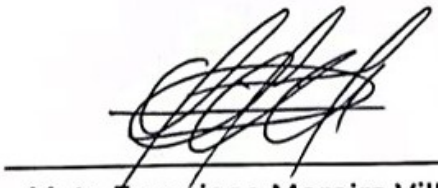


KEVIN QUELAL R.

TRIBUNAL DE GRADUACIÓN

Andrea García Angulo, Ph.D.
DIRECTOR

Purificación Galindo Villardon, Ph.D.
PRESIDENTE



Mgtr. Francisco Moreira Villegas
EVALUADOR

RESUMEN

En el contexto actual de transformación digital, las empresas del sector de consumo masivo están experimentando un cambio paradigmático hacia la utilización de datos para la toma de decisiones estratégicas. Se introduce una herramienta analítica basada en métodos multivariantes y aprendizaje estadístico para la segmentación eficiente de clientes y la optimización de portafolios de productos. Mediante el uso de algoritmos de clustering y técnicas de minería de datos como FP-Growth para el análisis de reglas de asociación, esta investigación propone una metodología para generar recomendaciones de compra personalizadas, buscando mejorar la eficacia de la fuerza de ventas y fortalecer la lealtad del cliente. El objetivo es evaluar la viabilidad de clasificar a los clientes según perfiles de compra específicos y desarrollar canastas de compras personalizadas que respondan a sus necesidades únicas. Este proyecto detalla la aplicación de estas técnicas para analizar datos de ventas y clientes de 2021, con el fin de ofrecer estrategias de marketing más dirigidas y efectivas en el competitivo mercado de consumo masivo.

Palabras Clave: Segmentación de Clientes, Métodos Multivariantes, Clustering, Reglas de Asociación, FP-Growth, Canastas de Compra Personalizadas.

ABSTRACT

In the current context of digital transformation, mass consumption sector companies are experiencing a paradigm shift towards the use of data for strategic decision-making. An analytical tool based on multivariate methods and statistical learning is introduced for efficient customer segmentation and product portfolio optimization. Through the use of clustering algorithms and data mining techniques like FP-Growth for association rule analysis, this research proposes a methodology for generating personalized purchase recommendations, aiming to improve the effectiveness of the sales force and strengthen customer loyalty. The goal is to assess the feasibility of classifying customers according to specific purchase profiles and developing personalized shopping baskets that meet their unique needs. This project details the application of these techniques to analyze sales and customer data from 2021, in order to offer more targeted and effective marketing strategies in the competitive mass consumption market.

Keywords: *Customer Segmentation, Multivariate Methods, Clustering, Association Rules, FP-Growth, Personalized Shopping Baskets.*

ÍNDICE GENERAL

ABREVIATURAS

ESPOL	Escuela Superior Politécnica del Litoral
SKU	Stock Keeping Unit
BIC	Bayesian information criterion
GMM	Gaussian Mixture Model
DBSCAN	Density-based Spatial Clustering of Applications with Noise
FP-Tree	Frequent Pattern Tree
FP GROWHT	Frequent Pattern growth

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

1.1 Antededentes

Durante los últimos años las empresas del sector de consumo masivo han tenido una evolución a nivel tecnológico y análisis de datos. La toma de decisiones organizacionales se ha convertido en un tema de investigación emergente en las últimas décadas ya que las decisiones estratégicas son la preocupación central para las organizaciones modernas (Rodríguez-Cruz and Pinto, 2018).

Se ha reconocido de manera generalizada que la toma de decisiones es fundamental para el desarrollo y éxito de las empresas, constituyendo una actividad organizacional primordial. En el pasado, este proceso se gestionaba de manera subjetiva, sin una consideración adecuada del entorno empresarial o la complejidad de factores que influyen en los procesos de ventas. Esta perspectiva ha evolucionado hacia un enfoque más analítico y basado en datos, reconociendo la importancia de integrar múltiples variables y tendencias del mercado en la toma de decisiones estratégicas (Olson, 2010).

En el dinámico entorno empresarial actual, las organizaciones generan una cantidad masiva de datos, lo cual plantea tanto un desafío como una oportunidad. Para capitalizar esta vasta cantidad de información, es esencial que los datos no solo sean recopilados, sino también meticulosamente procesados y analizados. Este proceso de transformación de grandes volúmenes de datos brutos en información útil y relevante se ha convertido en un pilar fundamental en la toma de decisiones estratégicas. Al implementar sistemas de análisis de datos avanzados, las empresas pueden identificar tendencias ocultas, prever cambios en el mercado, y tomar decisiones basadas en perspectivas accionables. Este enfoque basado en datos no solo mejora la precisión de las decisiones empresariales, sino que también permite una respuesta más rápida y efectiva a las condiciones cambiantes del mercado, asegurando que la empresa se

mantenga competitiva y relevante en su sector.

En las compañías dedicadas al sector de consumo masivo, su fuerza de ventas se enfrenta a determinadas retos. La eficacia de ellos es vital para el éxito y la sostenibilidad de las compañías en el sector de consumo masivo. Estos equipos, al estar en la primera línea de interacción con los clientes, desempeñan un papel crucial no solo en la generación de ingresos, sino también en la construcción y el mantenimiento de relaciones a largo plazo con los clientes. En este contexto, es imperativo que las decisiones tomadas por la fuerza de ventas no se basen únicamente en la intuición o en prácticas convencionales, sino que estén respaldadas por un análisis riguroso de la información de los clientes. Al aprovechar los datos para comprender mejor las preferencias y comportamientos de los clientes, las empresas pueden personalizar su enfoque de ventas, ofreciendo los productos más adecuados a cada cliente. Esta estrategia basada en datos no solo optimiza las oportunidades de venta, sino que también refuerza la lealtad del cliente y mejora la posición competitiva de la empresa en un mercado cada vez más saturado y competitivo. Entre los grandes retos que conlleva vender tenemos la presentación de un amplio portafolio de productos a los clientes. Este proceso suele ser complejo porque debe responder a la siguiente pregunta ¿Qué productos se deben ofertar a determinados clientes?

1.2 Justificación del problema

El creciente aumento de datos en las empresas, ha transformado la manera de tomar decisiones, debido a esto se introduce una metodología llamada Data Driven, la cual se fundamenta en la planificación estratégica, recopilación y análisis de información.(Seetharaman et al., 2016).

La empresa de consumo masivo, se dedica a ofertar productos a diferentes tipos de clientes, mediante su fuerza de ventas, los cuales visitan a los clientes asignados según su rutero de visita.

Existe un amplio portafolio por ofertar, dado esto no se podría ofertar el 100% de los productos, ya que los clientes tienen restricciones presupuestarias y la empresa no es la única que ofrece productos a los clientes, aquí parte la problemática del presente trabajo.

En las empresas de consumo masivo, una buena segmentación de clientes y productos puede llevar a ahorros significativos en descuentos y promociones. Este enfoque permite a las empresas dirigirse de manera más efectiva a los segmentos de mercado adecuados y optimizar sus estrategias de marketing para obtener el máximo impacto.

Los avances tecnológicos influyen directamente en el estilo de vida de las personas. Por esta razón, las empresas tradicionales deben adaptarse a esta evolución para seguir siendo relevantes en un nuevo contexto, ante nuevos gustos y nuevas necesidades. Ante esta realidad cambiante, descubrir los hábitos de consumo es la pregunta y aprendizaje estadístico es la respuesta tecnológica para acercarse a los compradores de hoy. Es de vital importancia buscar el portafolio adecuado para un grupo determinado de clientes con características similares. El aprendizaje estadístico y el análisis multivariante juegan un rol importante para que las empresas puedan anticiparse a estos escenarios realistas y buscar soluciones mediante el uso de herramientas tecnológicas y de análisis de datos.

Por tal motivo se quiere crear una herramienta que permita la caracterización de los clientes y la generación de sugerencias de mix de productos con una visualización dinámica e interactiva.

1.3 Objetivos

1.3.1 Objetivo General

Evaluar a través de métodos multivariantes la clasificación de clientes de una empresa del sector de consumo masivo de Guayaquil con el fin de generar una sugerencia de compra de productos que ayude a la elección del portafolio adecuado al cliente correcto.

1.3.2 Objetivos Específicos

1. Clasificar a los clientes de acuerdo a su perfil de compras y demás variables relevantes mediante la aplicación de metodología multivariante de segmentación.
2. Definir el portafolio de productos a los clientes mediante el uso de software especializado de aprendizaje estadístico.
3. Mostrar los principales hallazgos mediante el uso de herramientas tecnológicas y visualización dinámica e interactiva de los datos.

1.3.3 Hipótesis

La implementación de algoritmos de clustering, adaptados para considerar variables geoespaciales y otras características pertinentes, proveerá una segmentación de clientes más precisa para empresas de consumo masivo en Guayaquil. Esta segmentación avanzada es crucial para el desarrollo de canastas de compras personalizadas mediante algoritmos de reglas de asociación.

1.3.4 Alcance

El presente trabajo de investigación está enfocado en la búsqueda de métodos multivariantes para la caracterización de clientes. Asimismo, generar sugerencias de compra de productos a clientes de una empresa del sector de consumo masivo radicada en la ciudad de Guayaquil mediante aprendizaje estadístico. Para el presente trabajo de investigación se utilizarán datos oficiales de ventas y maestro de clientes domiciliados en la ciudad de Guayaquil de una empresa del sector de consumo masivo en el año 2021.

CAPÍTULO 2

2. MARCO TEÓRICO

Según Larose and Larose (2014), el auge del uso de los datos en las empresas ha ido creciendo a lo largo de los años, por lo que la información se ha convertido en un recurso vital para el desarrollo y evolución de cualquier empresa, en donde la competitividad hace necesaria la obtención de información de una manera rápida y eficiente frente al crecimiento diario de la misma.

Las ventas juegan un papel importante en cualquier empresa y las tareas realizadas por los vendedores son parte fundamental del crecimiento de los negocios. En base a la experiencia que tiene el cliente con los vendedores comprara uno o más de los productos ofertados, esto indica básicamente cuánto los clientes confían en los productos y en el negocio mismo, debido a esto podrían recomendar esos productos a las personas en su red (Walker and Mcclellandt, 1991).

Afortunadamente, las industrias son conscientes de la relevancia de las ventas y los datos de ventas y cómo pueden impulsar una variedad de decisiones, por lo que es crucial ayudar en los procesos comerciales y derivar mejores resultados. Además, las buenas ventas por sí solas no son suficientes y las fluctuaciones de las ventas a lo largo del tiempo es un problema importante que enfrentan la mayoría de las industrias por eso que los gerentes y los tomadores de decisiones buscan buenos modelos de predicción de ventas a través de los cuales la estabilidad pueda ser logrado (Das and Chaudhury, 2007).

La predicción de ventas es ahora una tarea no subestimada que cuenta con un equipo dedicado de estadísticos, analistas de datos y científicos. Por ejemplo, se utiliza para planificar ventas y logística de distribución y en economía, planificación estratégica. para la asignación de recursos se puede mejorar utilizando un buen pronóstico de ventas (Das and Chaudhury, 2007).

A pesar de su relevancia, la predicción de ventas es un proceso complejo y tiene

siempre ha sido un desafío porque depende de factores internos y externos. Variaciones en los gustos y demandas de los clientes, el estado financiero y el presupuesto de marketing establecido por un determinado supermercado se encuentran entre los internos. También hay algunos factores externos que influyen tales como la fluctuación general en la economía, las tendencias internacionales en los alimentos y la dieta hábitos y ocasiones especiales (Lu et al., 2014).

2.1 Segmetacion de clientes en las empresas

En el ámbito de la segmentación de clientes para empresas de consumo masivo, la modelización de clústeres se ha establecido como una técnica crucial debido a su eficacia en la identificación de patrones de comportamiento del consumidor y la optimización de estrategias de marketing. Investigadores como Martínez and Fernández (2021) argumentan que la agrupación de clientes en clústeres según características similares permite a las empresas diseñar estrategias de marketing más personalizadas y eficientes. Esta técnica es particularmente valiosa en mercados altamente saturados, donde entender las sutilezas y preferencias de los consumidores es clave para el éxito comercial. Se destaca la modelización de clústeres, especialmente cuando se combina con algoritmos de aprendizaje automático, puede revelar segmentos de clientes ocultos, ofreciendo oportunidades para innovar y capturar nichos de mercado previamente desatendidos. Estos estudios subrayan no solo la relevancia de la segmentación de clientes para las empresas de consumo masivo, sino también la creciente importancia de adoptar enfoques analíticos avanzados para mantener la competitividad en un mercado en constante cambio (Zhou and Wang, 2022).

2.2 Métodos de Clúster

Existen varios métodos estadísticos que se pueden utilizar para la segmentación de clientes. A continuación se presenta una revisión de tres de los más utilizados.

2.2.1 K Means

El algoritmo K-means, introducido por MacQueen (1967), es un método de análisis de clústeres que se ha establecido como uno de los enfoques más utilizados en el campo

del aprendizaje automático y la minería de datos. Este algoritmo tiene como objetivo particionar un conjunto de datos en K clústeres distintos, asignando cada punto de datos al clúster cuyo centroide (el promedio de todos los puntos en el clúster) es el más cercano.

Funcionamiento del Algoritmo

El proceso de K-means comienza con la selección aleatoria de K centroides, seguido de la asignación de cada punto de datos al clúster más cercano y la actualización de los centroides de cada clúster. Esta iteración continúa hasta que los centroides no cambian significativamente entre iteraciones sucesivas o se alcanza un número predeterminado de iteraciones (MacQueen, 1967).

Selección del Número de Clústeres

Determinar el número adecuado de clústeres (K) es una decisión crucial en el algoritmo K-means. Métodos como el ”método del codo” y el índice de Silhouette, introducido por (Rousseeuw, 1987), se utilizan para estimar el número óptimo de clústeres basándose en criterios de cohesión y separación de los clústeres.

Desafíos y Limitaciones

A pesar de su popularidad, el algoritmo K-means tiene varias limitaciones. Es sensible a los valores iniciales de los centroides y puede quedar atrapado en mínimos locales. Además, el algoritmo asume que los clústeres son esféricos y de tamaño similar, lo que puede no ser adecuado para algunos conjuntos de datos. Para abordar algunos de estos problemas, se han propuesto variantes como K-means++ para una inicialización más efectiva de los centroides (Arthur and Vassilvitskii, 2007).

Aplicaciones del K-means

El K-means se ha aplicado en numerosos campos, desde el análisis de mercado hasta el procesamiento de imágenes y la genómica. Su simplicidad y eficiencia lo hacen ideal para el análisis exploratorio y la identificación de patrones en grandes conjuntos de datos.

2.2.1.1 Medidas de Distancia

Todos los métodos de clustering tienen una cosa en común, para poder llevar a cabo las agrupaciones necesitan definir y cuantificar la similitud entre las observaciones. El término distancia se emplea dentro del contexto del clustering como cuantificación de la

similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia. La característica que hace del clustering un método adaptable a escenarios muy diversos es que puede emplear cualquier tipo de distancia, lo que permite al investigador escoger la más adecuada para el estudio en cuestión. A continuación, se describen una de las más utilizadas,la distancia euclidiana (Ultsch and Lötsch, 2022).

2.2.1.2 Distancia Euclidiana

La distancia euclidiana entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclidiana se calcula empleando el teorema de Pitágoras. Por ejemplo, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x, y), la distancia euclídea entre p y q viene dada por la ecuación:

$$d_{\text{euc}}(p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

Esta ecuación puede generalizarse para un espacio euclídeo n-dimensional donde cada punto está definido por un vector de n coordenadas:

$$p = (p_1, p_2, p_3, \ldots, p_n) \quad \text{y} \quad q = (q_1, q_2, q_3, \ldots, q_n)$$

$$\begin{aligned} d_{\text{euc}}(p, q) &= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \end{aligned}$$

Una forma de dar mayor peso a aquellas observaciones que están más alejadas es emplear la distancia euclídea al cuadrado. En el caso del clustering, donde se busca agrupar observaciones que minimicen la distancia, esto se traduce en una mayor influencia de aquellas observaciones que están más distantes (Ultsch and Lötsch, 2022).

2.2.2 Gaussian Mixture Model

El Modelo de Mezcla Gaussiana (GMM), una ecuación fundamental es la que define la probabilidad de que un punto de datos x provenga de una mezcla de K distribuciones

gaussianas:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

donde:

- $p(x)$ es la probabilidad de observar el punto de datos x bajo el modelo.
- K es el número de componentes gaussianos en el modelo.
- π_k son los coeficientes de la mezcla que satisfacen $\sum_{k=1}^{K} \pi_k = 1$ y $\pi_k \geqslant 0$, representando la probabilidad a priori de que x provenga del k-ésimo componente.
- $N(x|\mu_k, \Sigma_k)$ es la densidad de probabilidad de la distribución gaussiana para el k-ésimo componente, con media μ_k y covarianza Σ_k.
- μ_k es el vector de medias del k-ésimo componente gaussiano.
- Σ_k es la matriz de covarianza del k-ésimo componente gaussiano que determina su forma y orientación.

La estimación de los parámetros π_k, μ_k, y Σ_k se realiza a través del algoritmo Expectation-Maximization (EM), que itera entre los pasos de Expectation (E) y Maximization (M), comenzando con una suposición inicial de los parámetros y mejorándolos iterativamente para maximizar la verosimilitud del modelo dado los datos (Dempster et al., 1977).

El modelo de mezcla gaussiana (*Gaussian Mixture Model*, GMM) es un método probabilístico para la representación de subpoblaciones dentro de una población global, ampliamente utilizado en estadísticas y aprendizaje automático para tareas de agrupamiento (*clustering*) y clasificación (Bishop, 2006). Los GMM son una forma de modelado de mezcla que asume que todas las muestras de datos provienen de una combinación de varias distribuciones gaussianas desconocidas, con sus propios parámetros de media y varianza (McLachlan and Peel, 2000).

Cada componente del GMM representa una distribución gaussiana; esto es particularmente poderoso porque las distribuciones gaussianas pueden modelar una

amplia variedad de formas en el espacio de datos, lo que permite que los GMM capturen la complejidad en los datos multidimensionales (Reynolds, 2015). Además, a diferencia de los modelos de clustering más simples, como k-means, los GMM proporcionan no solo una asignación de cada observación a un *cluster*, sino también probabilidades de pertenencia, ofreciendo así una medida de incertidumbre en la clasificación de los puntos de datos (Rasmussen, 2000).

El enfoque de los GMM es particularmente útil en situaciones donde la estructura de los datos es desconocida o cuando los datos no están claramente separados. A través del uso del algoritmo *Expectation-Maximization* (EM), los GMM iterativamente ajustan los parámetros del modelo para maximizar la probabilidad de los datos, dado el modelo (Dempster et al., 1977). Este enfoque permite que los GMM converjan a soluciones que son óptimas localmente, aunque no hay garantía de encontrar un óptimo global, debido a la posible presencia de múltiples máximos locales en la función de probabilidad (Bilmes, 1998).

GMM es una herramienta flexible y poderosa para el análisis de datos, capaz de modelar complejas distribuciones subyacentes y proporcionar una rica comprensión de las estructuras de datos. Su aplicación se extiende a través de campos como el reconocimiento de patrones, la visión por computadora, y la bioinformática, demostrando su versatilidad y robustez como modelo estadístico (Fraley and Raftery, 2002).

2.2.3 Density based clustering (DBSCAN)

Density-based spatial clustering of applications with noise (DBSCAN) fue presentado en 1996 por Ester et al. (1996) donde el define este algoritmos como una forma de identificar clusters siguiendo el modo intuitivo en el que lo hace el cerebro humano, identificando regiones con alta densidad de observaciones separadas por regiones de baja densidad. Según Jang and Jiang (2018), los métodos de *partitioning clustering* como k-means, hierarchical, k-medoids, c-means, etc., son buenos encontrando agrupaciones con forma esférica o convexa que no contengan un exceso de *outliers* o ruido, pero fallan al tratar de identificar formas arbitrarias. De ahí que el único cluster que se corresponde con un grupo real son los que provienen de un algoritmo DBSCAN.

DBSCAN evita este problema siguiendo la idea de que, para que una observación forme parte de un cluster, tiene que haber un mínimo de observaciones vecinas dentro de un radio de proximidad y de que los clusters están separados por regiones vacías o con pocas observaciones.
El algoritmo DBSCAN necesita dos parámetros:

- **Epsilon** (ϵ): radio que define la región vecina a una observación, también llamada (ϵ)-neighborhood.
- **Minimum points** (*minPts*): número mínimo de observaciones dentro de la región epsilon.

Por último, empleando las tres categorías anteriores se pueden definir tres niveles de conectividad entre observaciones:

1. **Directamente alcanzable** (*direct density reachable*): una observación A es directamente alcanzable desde otra observación B si A forma parte del (ϵ)-neighborhood de B y B es un core point. Por definición, las observaciones solo pueden ser directamente alcanzables desde un core point.
2. **Alcanzable** (*density reachable*): una observación A es alcanzable desde otra observación B si existe una secuencia de core points que van desde B a A.
3. **Densamente conectadas** (*density connected*): dos observaciones A y B están densamente conectadas si existe una observación core point C tal que A y B son alcanzables desde C.

2.3 Algoritmos de minería de reglas de asociación

Para explorar patrones en transacciones de productos a gran escala, se emplean reglas de asociación. Estas reglas ayudan a identificar relaciones habituales entre productos. Para manejar el análisis extenso de estas reglas, se recurre a técnicas de aprendizaje automático, destacando entre ellas el uso de algoritmos como A priori y FP-Growth. Estos métodos facilitan el procesamiento efectivo y la extracción de insights valiosos de grandes conjuntos de datos transaccionales.
A continuación se realiza una revisión de cómo funcionanestos algoritmos.

2.3.1 Algoritmo A priori

El algoritmo Apriori es un método clásico y fundamental en la minería de reglas de asociación. El algoritmo se basa en el principio de que un subconjunto de un conjunto de ítems frecuente también debe ser frecuente, una propiedad conocida como anti-monotonía del soporte . Esta propiedad es esencial para el proceso del algoritmo, ya que reduce significativamente el espacio de búsqueda de conjuntos de ítems frecuentes (Agrawal and Srikant, 1994).

El Apriori opera en un enfoque de abajo hacia arriba, comenzando con el cálculo de conjuntos frecuentes de un solo ítem y expandiéndose a conjuntos más grandes con cada iteración. En cada etapa, los conjuntos de ítems candidatos se generan utilizando los conjuntos frecuentes identificados en la iteración anterior. Los candidatos que no alcanzan el soporte mínimo son podados, y no se generan sus superconjuntos en las siguientes iteraciones (Agrawal and Srikant, 1994).

El soporte se define como la proporción de transacciones que contienen un conjunto de ítems, y la confianza mide la probabilidad condicional de encontrar un conjunto de ítems Y dado otro conjunto X, para una regla de asociación $X \Rightarrow Y$ (Savasere et al., 1995).

2.3.2 Algoritmo FP-Growth

El algoritmo FP-Growth, representa un avance significativo en la minería de reglas de asociación, especialmente en contextos donde se manejan grandes bases de datos. Este algoritmo se diferencia del Apriori al adoptar un enfoque más eficiente que evita la generación exhaustiva de candidatos. Utiliza una estructura de datos compacta, el FP-Tree (Frequent Pattern Tree), para almacenar información clave de conjuntos de ítems frecuentes de manera eficiente (Han et al., 2000).

La construcción del FP-Tree es un proceso de dos etapas: inicialmente, se realiza un escaneo de la base de datos para identificar los ítems frecuentes, seguido por la construcción del árbol en una segunda pasada, utilizando estos ítems ordenados por frecuencia. Esta metodología reduce significativamente la necesidad de múltiples exploraciones de la base de datos, una limitación principal del método A priori (Han et al., 2000).

En la fase de extracción de conjuntos de ítems frecuentes, el algoritmo FP-Growth explota un método de ”crecimiento de patrones”, dividiendo el árbol en subárboles y combinando patrones de prefijos comunes. Este enfoque evita la generación de conjuntos candidatos y mejora la eficiencia y escalabilidad, sobre todo en bases de datos de gran tamaño (Han et al., 2000).

La relevancia del algoritmo FP-Growth en la minería de datos es notable, siendo ampliamente utilizado para descubrir patrones y asociaciones en grandes conjuntos de datos, donde otros métodos como el Apriori resultan imprácticos debido a su alta demanda computacional (Han et al., 2000).

CAPÍTULO 3

3. METODOLOGÍA

3.1 Datos

La información relacionada a las variables de ventas se la recopiló de una fuente primaria interna proporcionada por la Empresa XYZ perteneciente al sector consumo masivo de la ciudad de Guayaquil.

3.1.1 Comprensión de los datos del negocio

La estructura del sistema de información comercial de ventas del negocio se encuentra definida a partir del modelo dimensional.Como se puede apreciar en la figura 3.1, en el sistema de negocio de la Empresa XYZ, el modelo dimensional de ventas se centra en una tabla de hechos que representa las ventas realizadas. Esta tabla está asociada con dos dimensiones principales: clientes y productos. Cada registro en este modelo representa un único ítem de factura asociado a un producto específico, vendido por la Empresa XYZ a un cliente en particular durante el año 2021. La granularidad de los datos en este modelo es tal que cada transacción de venta de producto a cliente se registra de manera individual es decir con un registro diario.

Figura 3.1. Modelo dimensional de la Empresa XYZ

3.1.2 Tablas

Detalle de las tablas involucradas en el modelo dimensional de ventas

Tabla 3.1. Diseño de la tabla de hechos ventas

Nombre del atributo	Tipo de datos
IDCodigoCliente	integer
VolumenTotal	numeric
Cajas	numeric
Descuento	integer
VentaBruta	integer
VentaNeta	integer
IDSKUfinal	character

Tabla 3.2. Diseño de la dimensión cliente

Nombre del atributo	Tipo de datos
IDCodigoCliente	integer
latitud	integer
longitud	integer
estado cliente	character
IDVendedor	character

Tabla 3.3. Diseño de la dimensión producto

Nombre del atributo	Tipo de datos
IDLinea	character
IDSKUfinal	character

3.1.3 Procesamiento de datos

En esta fase del estudio, el primer paso consistió en adquirir las bases de datos transaccionales de la empresa en cuestión. Se logró obtener acceso a tres conjuntos de datos clave: la tabla de ventas correspondiente al año 2021, el registro de clientes históricos ubicados en la ciudad de Guayaquil, y la base de datos de los productos de la empresa. La base de ventas tiene una dimensionalidad de 1.4 millones de registros, la

base de clientes de 22.6 mil registros y finalmente la base de productos de 638 productos.

El análisis exploratorio y análisis de información se lo realizó en Python 3.9.13, este permitió el tratamiento de la información, exploración de datos y para finalmente la construcción de los algoritmos de clusterización y asociación de productos.

3.2 Modelo de Clusterización

La segmentación de clientes con técnicas de clustering ofrece una base sólida para la toma de decisiones basada en datos, lo que resulta en estrategias más informadas y efectivas (Wedel and Kamakura, 2000).

En esta parte se despliega la metodología adoptada para la exploración y análisis de clusters. Los algoritmos de clustering seleccionados para este trabajo fueron: K Means, Modelos de Mezclas Gaussianas (GMM) y DBSCAN (Densidad de Conexión Basada en Algoritmo de Agrupación Espacial de Aplicaciones con Ruido), son examinados no solo por su aplicabilidad práctica sino también por su relevancia teórica y su capacidad para proporcionar perspectivas distintivas en la estructura inherente de los conjuntos de datos.

3.2.1 Definiciones de variables para clusterización

Se adoptó una metodología cuantitativa para analizar patrones subyacentes en los datos de los clientes. La metodología se centra en técnicas avanzadas de análisis de datos, incluyendo preprocesamiento, análisis de componentes principales (PCA), clustering y evaluación cuantitativa de la calidad de los clusters. Se detalla cada etapa del proceso analítico.

Las variables utilizadas para la clusterización son: IDCodigoCliente, canal, latitud, longitud junto con el promedio del volumen total de transacciones del año 2021.

3.2.2 Preprocesamiento de datos

Se implementó un preprocesamiento robusto para transformar los datos originales en un formato adecuado para el análisis. Esto incluyó la agrupación por cliente y canal, y la agregación del volumen total de transacciones.

3.2.2.1 Transformación de coordenadas

Posteriormente se aplicó un tratamiento de las coordenadas geográficas (latitud y longitud) en su forma original para análisis estadísticos, incluido el clustering, presenta desafíos. Estos desafíos se deben a la naturaleza cíclica y esférica de las coordenadas, especialmente cuando se abordan a escala global (Fotheringham et al., 2000). La transformación de estas coordenadas en valores de seno y coseno mitiga estos problemas, permitiendo un análisis más robusto y significativo.

La conversión de latitud y longitud a sus componentes de seno y coseno se basa en principios matemáticos de trigonometría esférica. Esta transformación se realiza para abordar la circularidad de los datos, permitiendo que las técnicas de análisis de clustering operen en un espacio más 'euclidiano', facilitando así la identificación de clusters (Banerjee et al., 2004).

3.2.3 Normalización

Los datos se estandarizaron utilizando StandardScaler de la biblioteca sklearn, lo cual es crucial para técnicas como PCA y modelos de clustering que son sensibles a la escala de las características.

3.2.4 Reducción de dimensionalidad mediante PCA

El PCA se utilizó para reducir la dimensionalidad del conjunto de datos, facilitando la visualización e interpretación de los datos, y eliminando posibles correlaciones entre variables.

Después de aplicar PCA, analizamos los pesosde los componentes principales para entender qué variables contribuyen más a cada componente. Utilizamos un mapa de calor para visualizar estas relaciones.

3.2.5 Clustering

3.2.5.1 Clustering mediante K-Means

Empleamos el método Elbow (Agrawal and Srikant, 1994), utilizando KElbowVisualizer, para determinar el número óptimo de clústeres. Esta técnica evalúa la varianza explicada en función del número de clústeres y elige el punto donde se observa un

cambio significativo en la pendiente (elbow).

Implementamos K-Means con el número óptimo de clústeres. Es importante destacar que K-Means utiliza la distancia euclidiana como métrica por defecto. Esta elección es apropiada para nuestros datos normalizados y transformados, ya que la distancia euclidiana es efectiva en espacios de características homogéneos y bien escalados.

3.2.5.2 Clustering mediante GMM

Se exploró un enfoque alternativo de clustering utilizando el Modelo de Mezclas Gaussianas (GMM). GMM es un método probabilístico que asume que los datos se generan a partir de una mezcla de varias distribuciones gaussianas con parámetros desconocidos.

Se utilizo el Criterio de Información Bayesiana (BIC) para determinar el número óptimo de componentes gaussianos. El BIC proporciona un medio para evaluar el modelo considerando tanto la complejidad del modelo (número de componentes) como su ajuste a los datos. Se selecciono el número de componentes que minimiza el BIC.

Con el número óptimo de componentes determinado, se ajustó un modelo GMM a los datos. Este modelo no solo asigna cada punto a un clúster, sino que también proporciona probabilidades de pertenencia a cada clúster, lo que ofrece una visión más matizada que el enfoque más rígido del K-Means.

3.2.5.3 Clustering mediante DBSCAN

Además de **K-Means** y **GMM**, exploramos el clustering utilizando **DBSCAN** (*Density-Based Spatial Clustering of Applications with Noise*), un algoritmo basado en la densidad que puede identificar clústeres de formas arbitrarias y detectar puntos de ruido (*outliers*).

DBSCAN se caracteriza por sus parámetros *eps* (distancia máxima entre dos puntos para ser considerados vecinos) y *min_samples* (número mínimo de puntos para formar un clúster). Ajustamos estos parámetros de acuerdo con las características específicas de nuestro conjunto de datos.

Una ventaja clave de **DBSCAN** es su capacidad para identificar y excluir el ruido,

clasificando estos puntos con la etiqueta -1.

3.2.5.4 Evaluación de los clusters

Para evaluar la efectividad del clustering, utilizamos tres métricas distintas: el Silhouette Score, el Índice Davies-Bouldin y el Índice Calinski-Harabasz, cada una ofreciendo una perspectiva única en la calidad de los clústeres formados.

1. **Silhouette Score**:

   El *Silhouette Score* es una medida que evalúa cuán similares son los puntos dentro de un clúster en comparación con puntos en clústeres vecinos. Valores cercanos a $+1$ indican que los puntos están bien agrupados dentro de su propio clúster y lejos de otros clústeres (Rousseeuw, 1987).

2. **Índice Davies-Bouldin**:

   Este índice mide la calidad de los clústeres basándose en la relación entre la dispersión interna del clúster y la separación entre clústeres. Un valor bajo indica que los clústeres están densamente agrupados y bien separados (Davies and Bouldin, 1974).

3. **Índice Calinski-Harabasz**:

   Este índice compara la dispersión entre clústeres con la dispersión dentro de los clústeres. Valores altos indican clústeres bien definidos y separados (Caliński and Harabasz, 1974).

3.3 FP Growth

3.3.1 Preparación y Agrupación de Datos

Esta fase implica la preparación de los datos para el análisis de patrones frecuentes. Los datos se agrupan y se resumen para reflejar características clave como el ID del cliente, canal, SKU, mes y año. Este proceso ayuda a refinar el conjunto de datos, enfocándose en los aspectos más relevantes para el análisis de patrones de compra. La calidad y la

estructura adecuada de los datos son esenciales para cualquier análisis de minería de datos (Han et al., 2011).

3.3.2 Codificación y Preparación de Datos para FPGrowth

Los datos se transforman en un formato adecuado para el análisis mediante FPGrowth. Se utiliza el TransactionEncoder para convertir los datos a un formato binario, que es un paso esencial antes de aplicar algoritmos de patrones frecuentes. Esta codificación permite al algoritmo procesar eficientemente grandes conjuntos de datos (Borgelt, 2005).

3.3.3 Aplicación del Algoritmo FPGrowth

FPGrowth es un algoritmo conocido por su eficiencia en el procesamiento de grandes conjuntos de datos, lo que reduce significativamente la sobrecarga computacional. Este algoritmo es especialmente útil en bases de datos con una gran cantidad de transacciones o ítems (Han et al., 2011).

3.3.4 Generación y almacenamiento de reglas de asociación

Una vez identificados los conjuntos de ítems frecuentes, se utilizan para generar reglas de asociación. Estas reglas proporcionan insights sobre cómo los productos se compran juntos y pueden ser utilizados para informar estrategias de marketing y promoción. La confianza y otras métricas de interés se calculan para evaluar la fuerza y relevancia de estas reglas (Agrawal and Srikant, 1994).

3.3.5 Visualización y Evaluación de Resultados

Los resultados del análisis se visualizan para una interpretación más fácil. La visualización puede incluir gráficos que muestren las relaciones entre diferentes ítems y la fuerza de estas relaciones. La evaluación de los resultados se hace a través de métricas como el soporte, la confianza y el lift, proporcionando una comprensión profunda de la calidad y utilidad de las reglas generadas (Hahsler et al., 2005).

CAPÍTULO 4

4. RESULTADOS

4.1 Análisis descriptivo de la empresa

La figura 4.1 ilustra la distribución mensual de las ventas netas de una empresa o sector durante un año calendario. Se observa que las ventas netas se mantienen relativamente constantes, con valores que varían entre 7.5 millones y 9.6 millones en la mayoría de los meses. Sin embargo, se destaca un aumento significativo en el último mes, diciembre (Mes 12), donde las ventas netas ascienden a 13.6 millones, lo que sugiere una posible estacionalidad o un aumento de la demanda durante el período de festividades. Los meses de febrero, marzo y noviembre presentan las cifras más bajas, todas en 7.5 millones, mientras que los meses de abril y octubre muestran valores ligeramente superiores. Este patrón puede ser indicativo de ciclos de compra de los consumidores o de eventos específicos que afectan las ventas netas. Es importante considerar estos factores al planificar la producción, la gestión de inventario y las estrategias de marketing para alinearlos con las tendencias observadas en las ventas.

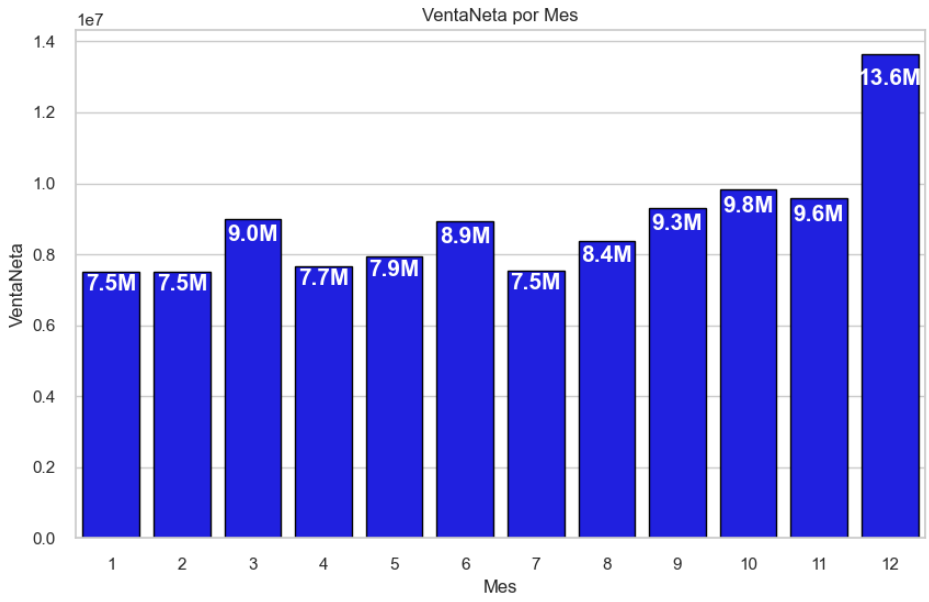


Figura 4.1. Venta Neta del año 2021 de la empresa XYZ

El figura 4.2 muestra los ingresos netos por canal de venta, con dos categorías: **TRADICIONAL** y **MAYORISTA**. En el canal **TRADICIONAL**, la mediana está muy cerca del borde inferior de la caja, lo que indica una concentración de valores de ingresos netos hacia el extremo inferior de la escala. Además, la caja, que representa el rango intercuartílico, es pequeña, sugiriendo que los valores están agrupados y hay poca variabilidad entre el primer y tercer cuartil.

En el caso del canal **MAYORISTA**, la mediana también se inclina hacia el borde inferior de la caja, pero esta es más larga que en el canal **TRADICIONAL**, lo que refleja una variabilidad mayor en los ingresos netos. Este canal muestra una cantidad significativa de valores atípicos, indicando la presencia de ventas con ingresos netos mucho mayores en comparación con el grueso de los datos.

Al comparar ambos canales, el **MAYORISTA** no solo presenta una variabilidad más alta, sino que también tiende a tener ingresos netos más elevados, como lo sugieren tanto la mediana como los valores atípicos. Estos valores atípicos indican transacciones que son excepcionalmente más altas que la mayoría de las ventas en ambos canales, pero son especialmente notorios y extremos en el canal **MAYORISTA**.

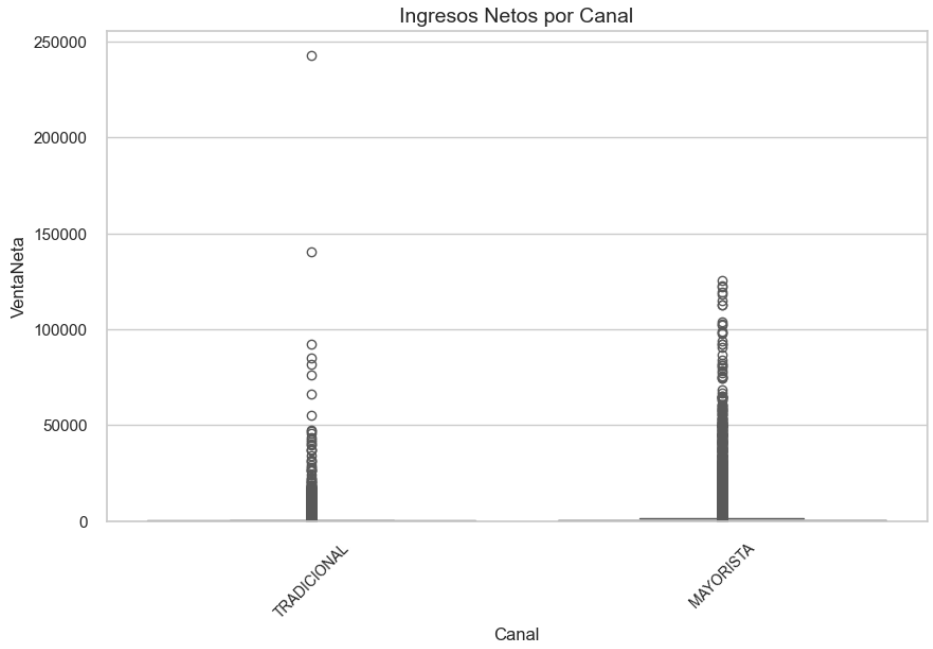


Figura 4.2. Distribución de Ingresos Netos entre Canales de Venta del año 2021 de la empresa XYZ

La figura 4.3 es un diagrama de barras apiladas que representa el volumen total de ventas mensuales de las tres principales unidades de mantenimiento de stock (SKUs), a lo largo de un año, dividido en 12 meses. Cada mes muestra tres barras que representan los tres principales SKUs, y cada barra está segmentada en colores que corresponden a diferentes *SKU*. Estos segmentos de colores representan la contribución de cada *SKU* al volumen total de ventas del SKU correspondiente.

Se observan cinco colores diferentes que representan cinco distintos *SKU*, y sus contribuciones específicas varían mes a mes. Los números en la parte superior de cada segmento representan la cantidad de unidades vendidas en miles para ese *SKU* específico en ese mes particular. En el mes 1 (enero), el SKU más a la izquierda tiene un volumen de ventas que consta de 11,000 unidades del *SKU* 320, 12,000 unidades del *SKU* 30, y 13,000 unidades del *SKU* 20.

El gráfico también muestra una tendencia de que algunos *SKU*, como el 20 y el 30, tienen una presencia constante en la mayoría de los meses, mientras que otros tienen una contribución más esporádica o en menor volumen. Además, se puede notar que en algunos meses, como el 12 (diciembre), hay un pico significativo en el volumen de

ventas con 15,000 unidades del *SKU* 20, lo que podría indicar una demanda estacional o una promoción exitosa.

El gráfico permite comparar el rendimiento de ventas de diferentes *SKU* a lo largo del año, proporcionando una visión clara de cuáles son los más vendidos y en qué momento.

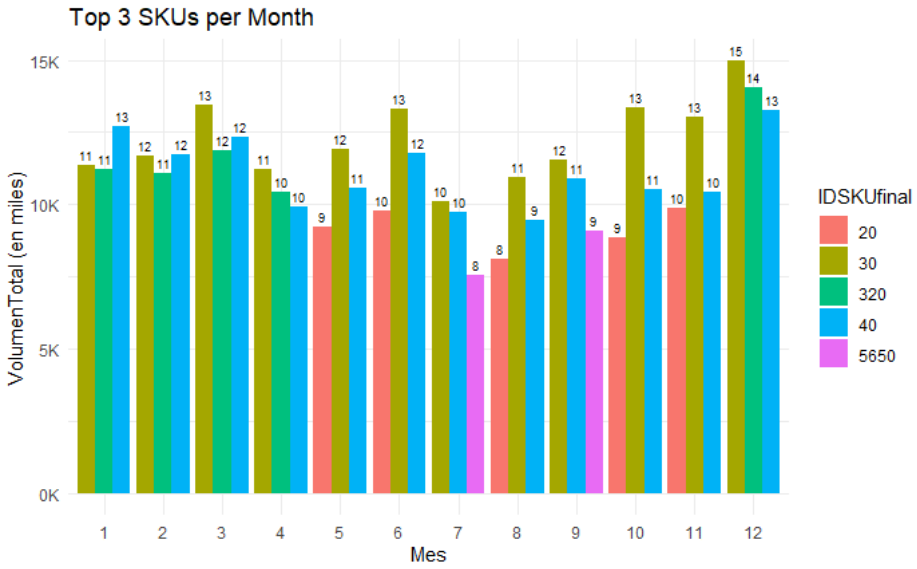


Figura 4.3. Volumen Total Mensual por SKU

La figura 4.4 muestra el número total de transacciones únicas de SKUs por mes, con los valores numéricos colocados estratégicamente para una lectura clara. La línea muestra las variaciones mensuales en el número de transacciones únicas de SKUs, lo que proporciona una visión clara de la actividad de ventas a lo largo del año 2021.

Podemos observar que el número de transacciones únicas de SKUs incrementa de enero a marzo, luego hay una caída en abril y mayo, y un pico significativo en junio, seguido de fluctuaciones en los meses subsiguientes. Hay una caída notable en octubre, seguida de un fuerte incremento hacia el final del año en noviembre y diciembre.

Este patrón podría reflejar la estacionalidad en las preferencias de compra o en la introducción de nuevos productos al mercado. Estos insights pueden ser valiosos para la planificación de estrategias de inventario y marketing, así como para la optimización del catálogo de productos.

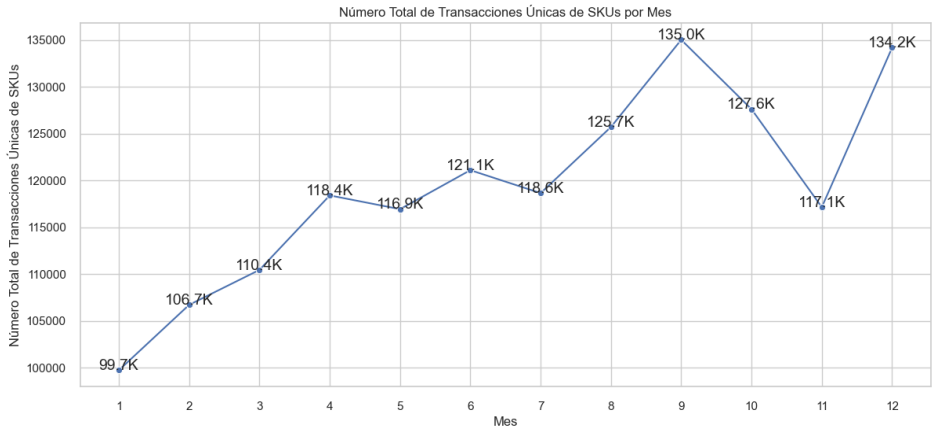


Figura 4.4. Tendencia Mensual del Número de Transacciones Únicas de SKUs

En el análisis de la figura 4.5 cantidad promedio de productos únicos (*IDSKUfinal*) adquiridos mensualmente, se observa una media aproximada de 8 productos únicos por cliente. La consistencia en la compra mensual se refleja en un rango de variabilidad relativamente estrecho, con un mínimo observado cerca de 7.1 y un máximo aproximado de 8.3 productos únicos. La desviación estándar visualmente estimada sugiere una baja variabilidad, indicando que los clientes tienden a comprar una cantidad similar de productos únicos mes a mes.

La interpretación de los cuartiles, aunque no numéricamente especificada, sugiere que el 50% de los datos se concentra en torno a la media con una variación modesta, lo que demuestra una estabilidad en las preferencias de compra de los clientes. Los valores de los cuartiles superior e inferior no muestran una dispersión significativa, lo que respalda la conclusión de que no existen fluctuaciones extremas en el comportamiento de compra a lo largo del año.

Este patrón de compra regular puede ser indicativo de una sólida fidelización de clientes o de una estrategia de inventario bien gestionada que mantiene un catálogo de productos coherente y alineado con las expectativas de los consumidores. Los insights derivados de este análisis podrían ser cruciales para la planificación de estrategias de marketing y ventas, asegurando la disponibilidad de productos y posiblemente aumentando la satisfacción del cliente.

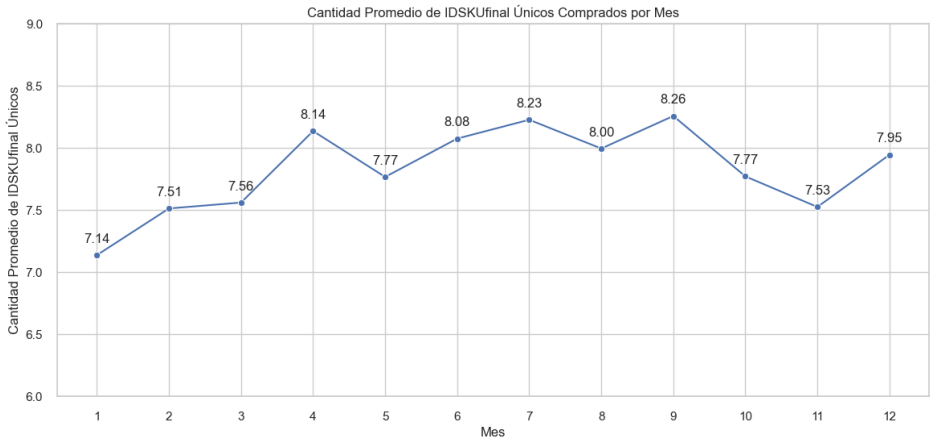


Figura 4.5. Tendencia Mensual del Número de Transacciones Únicas de SKUs

4.2 Segmentacion de los clientes mediante métodos de clusterizacion

4.2.1 PCA

La técnica de reducción de dimensionalidad, **PCA**, se aplicó al conjunto de datos escalado para identificar las direcciones principales que capturan la variabilidad más significativa dentro del conjunto de datos. Los resultados revelaron que el primer componente principal explica el 52.05% de la varianza, lo que refleja su papel predominante en la captura de la estructura subyacente de los datos. El segundo componente principal aporta una explicación adicional del 27.97% de la varianza. En conjunto, los dos primeros componentes principales representan aproximadamente el 80.02% de la variabilidad total, lo que indica que una proporción considerable de la información contenida en los datos originales puede ser resumida efectivamente en estas dos dimensiones.

4.2.2 Clusterización

Kmeans

Elección del número óptimo de clusters

Método del codo

La figura 4.6 representa el análisis del método del codo se aplicó para determinar el número óptimo de clústeres para el modelo de agrupación. La puntuación de distorsión

disminuyó rápidamente con el incremento de clústeres de 2 a 4, lo que indica una mejora sustancial en la cohesión de clúster. Sin embargo, al alcanzar $k = 5$, se observó un punto de inflexión, después del cual las ganancias marginales en la reducción de la distorsión se redujeron notablemente, señalando $k = 5$ como el número óptimo de clústeres con una puntuación de distorsión de aproximadamente 32132.763. Este punto de inflexión se identifica como el 'codo', más allá del cual la adición de clústeres adicionales no aporta una mejora significativa en la homogeneidad interna de los clústeres. El tiempo de ajuste del modelo mostró una leve tendencia ascendente a medida que k aumentaba, estabilizándose en clústeres más altos, lo que sugiere que el costo computacional adicional de más clústeres no es prohibitivo en este rango.

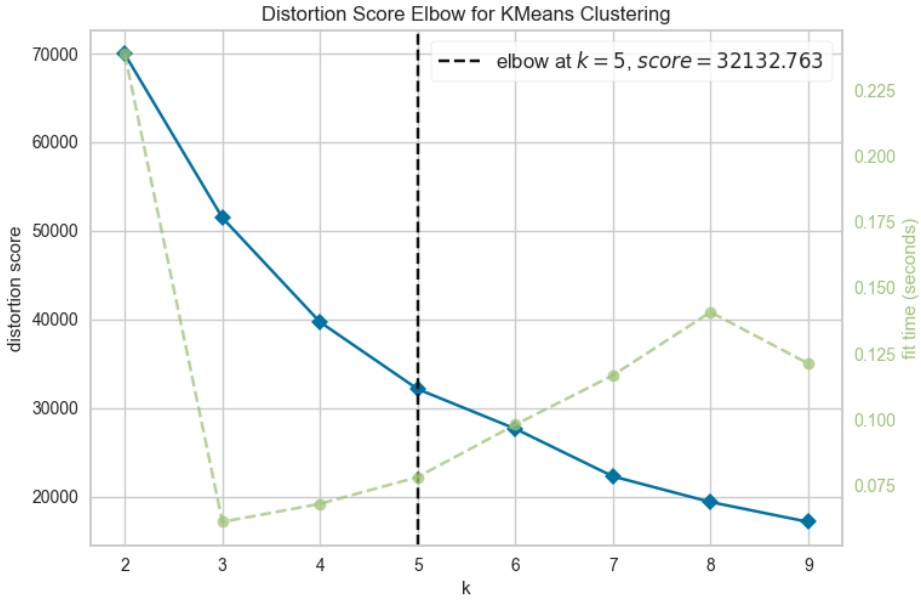


Figura 4.6. Análisis del Codo para la Selección de k en Clustering K-Means

En la figura 4.7 se ilustra la distribución de los datos tras la aplicación del algoritmo **K-Means** de clústeres sobre el conjunto de datos reducido mediante **PCA**. Se identificaron cinco clústeres distintos, cada uno visualizado con un color único, representando grupos diferenciados en el espacio de las dos componentes principales que resumen la variabilidad de los datos. El primer componente principal se extiende a lo largo del eje horizontal, mientras que el segundo componente principal se encuentra a lo largo del eje vertical. Los clústeres demuestran la heterogeneidad dentro del conjunto

de datos, con variaciones tanto en la densidad de los puntos como en su distribución espacial. Esta segmentación refleja la presencia de subgrupos dentro de los datos que comparten características similares, facilitando una comprensión más profunda de las estructuras intrínsecas del conjunto de datos analizado.

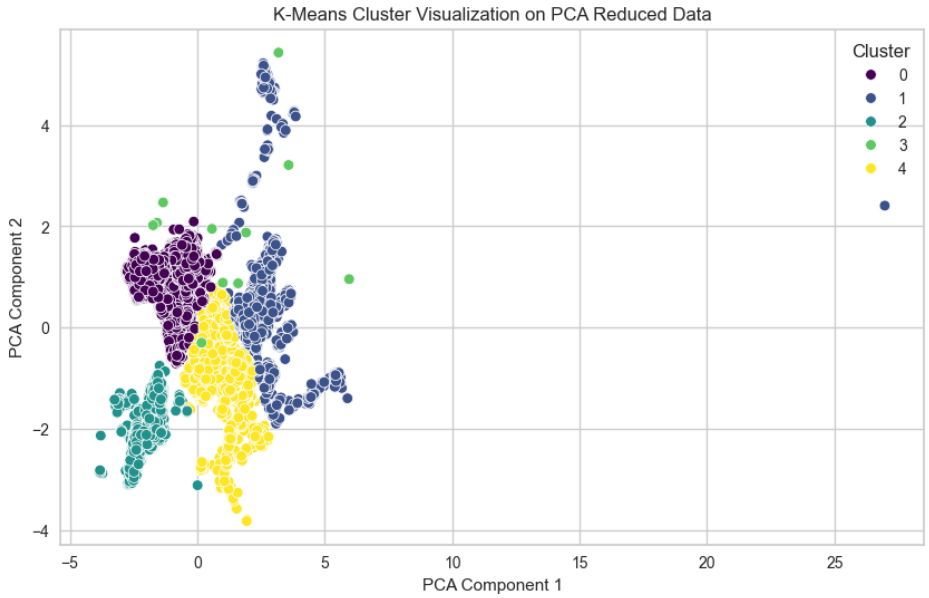


Figura 4.7. Distribución de Clusters con KMeans en Componentes Principales de Análisis PCA

En la tabla 4.1 se muestra la calidad del esquema de clústering resultante fue evaluada mediante tres métricas estadísticas consolidadas. El **Silhouette Score** obtenido fue de 0.4693, que, estando más cerca de 1 que de 0, sugiere una razonable separación y cohesión entre los clústeres; los puntos dentro de un clúster están más cerca entre sí en comparación con puntos en clústeres diferentes. No obstante, el valor indica que aunque hay una estructura de clústeres discernible, puede haber cierto solapamiento o ambigüedad en la asignación de algunos puntos.

El **Índice Davies-Bouldin**, con un valor de 0.7475, señala una distancia moderada entre los clústeres y densidades dentro de los clústeres relativamente similares. Valores más bajos de este índice indican una mejor separación de clúster, por lo que el valor obtenido refleja una calidad de clúster aceptable pero mejorable.

Finalmente, el **Índice Calinski-Harabasz** arrojó un valor de 14276.37, que es

considerablemente alto y, por lo tanto, denota una buena definición de clúster. Este índice mide la dispersión entre clústeres en comparación con la dispersión dentro de los clústeres, y un valor más alto es preferible, lo que implica que los clústeres están bien dispersos y bien separados.

En conjunto, estas métricas indican que el modelo de clústering ha identificado agrupaciones con una cohesión y separación adecuadas. El **Silhouette Score** positivo sugiere que los clústeres están bien definidos y separados en comparación con un escenario de asignación aleatoria. A pesar de que el valor no es cercano a 1, implica que hay una estructura de clúster significativa. El **Índice Davies-Bouldin**, aunque no óptimo, está por debajo de 1, lo que sugiere una separación razonable entre los clústeres. El alto valor del **Índice Calinski-Harabasz** confirma que la dispersión entre los clústeres es notable en comparación con la dispersión dentro de ellos. Sin embargo, dado que ninguna métrica proporciona una evidencia concluyente por sí sola, la combinación de estas tres sugiere que el esquema de clústering es bastante robusto pero podría beneficiarse de una afinación adicional para lograr una separación y cohesión óptimas entre los clústeres.

Tabla 4.1. Metricas Clusterización K Means

Métrica	Valor
Silhouette Score	0,4693
Índice Davies-Bouldin	0,7475
Índice Calinski-Harabasz	14276,37

Gaussian Mixture Model

Elección del número óptimo de clusters

La selección del número óptimo de clusters para la agrupación de datos mediante modelos de mezcla gaussiana fue guiada por el *Criterio de Información Bayesiano (BIC)*. Como se ilustra en la Figura 1, el análisis reveló una tendencia decreciente en los valores de BIC con un aumento en el número de clusters. Se observó una disminución pronunciada del BIC al pasar de un modelo de 1 cluster a uno de 2 clusters, lo cual indica una mejora significativa en la calidad del ajuste con la introducción de un segundo

cluster. Esta tendencia continuó a medida que el número de clusters aumentó, con una disminución más gradual entre los modelos de 2 a 9 clusters.

En la figura 4.8, el punto de inflexión se detectó en el modelo con 10 clusters, donde el valor de BIC alcanzó su mínimo. La presencia de una estrella magenta en el gráfico señala el número óptimo de clusters, que en este caso es 10, como se determinó por el valor más bajo de BIC. Este resultado sugiere que un modelo GMM con 10 clusters proporciona un compromiso óptimo entre la complejidad del modelo y la capacidad de ajustarse a la estructura subyacente de los datos. La estrella resaltada indica claramente este punto óptimo, facilitando la identificación visual del número de clusters recomendado para la segmentación de datos en el contexto de nuestro estudio.

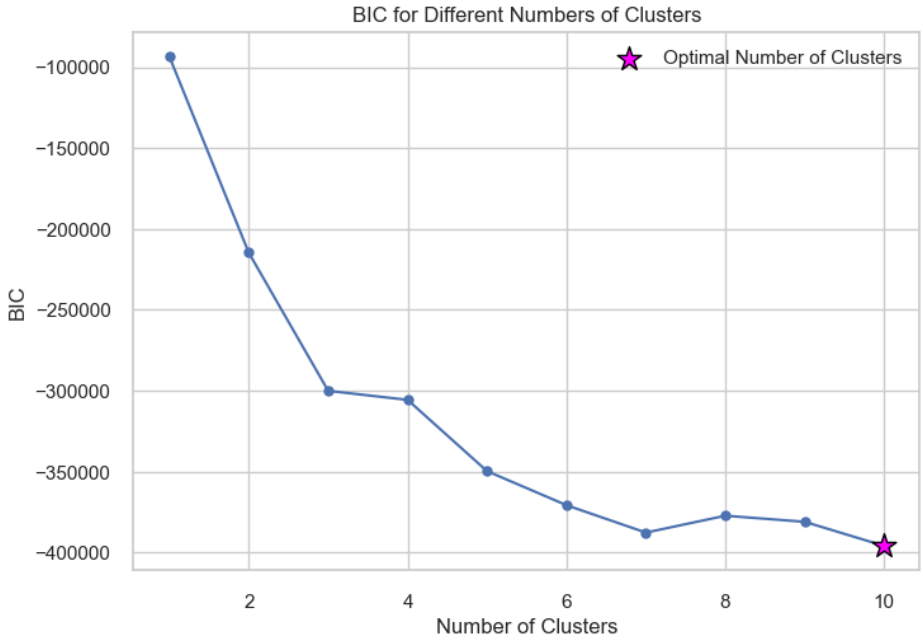


Figura 4.8. Selección del Número Óptimo de Clusters según el Criterio de Información Bayesiano (BIC)

La Figura 4.9 muestra la distribución de los clústeres obtenidos tras aplicar un modelo de mezcla gaussiana a nuestro conjunto de datos, el cual previamente había sido sometido a una reducción de dimensionalidad mediante Análisis de Componentes Principales. Se seleccionaron los dos componentes principales que capturan la mayor variabilidad de los datos para visualizar el resultado del clustering en un espacio bidimensional. Las observaciones se agrupan en 10 clústeres distintos, cada uno representado por un color

único en el gráfico.

La agrupación de los puntos en el espacio de las componentes principales indica patrones intrínsecos y relaciones subyacentes en los datos multidimensionales originales. Se observa una clara delimitación entre varios de los clústeres, lo que sugiere diferencias sustanciales en las características subyacentes que definen cada grupo. Algunos clústeres muestran una alta densidad y están bien definidos, como es el caso del clúster amarillo en la región inferior izquierda del gráfico, mientras que otros, como el clúster verde en la extrema derecha, consisten en pocas observaciones dispersas, lo que podría indicar datos atípicos o un comportamiento anómalo.

Esta representación gráfica facilita la comprensión de la estructura compleja de los datos y apoya la interpretación cualitativa de los resultados del clustering. La efectividad de la segmentación será posteriormente validada a través de métricas de evaluación de clustering, como el Silhouette Score, el Índice Davies-Bouldin y el Índice Calinski-Harabasz.

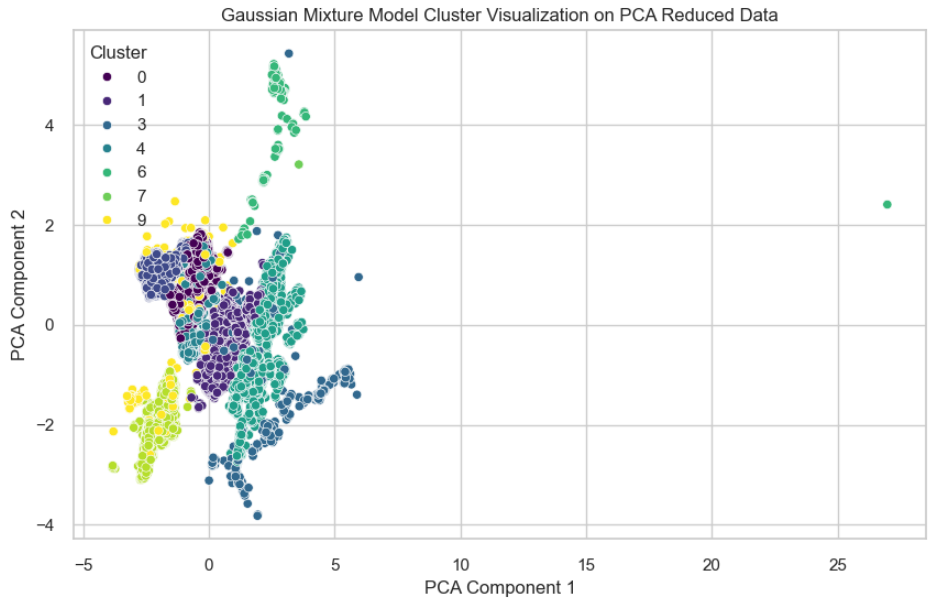


Figura 4.9. Distribución de Clusters con GMM en Componentes Principales de Análisis PCA

En la tabla 4.2 se muestran las métricas de evaluación de clustering utilizadas para cuantificar la calidad de los clústeres derivados del modelo GMM sobre los datos PCA-reducidos revelan una segmentación moderadamente efectiva. El **Silhouette**

Score obtenido fue de 0.3117, indicando una adecuada cohesión y separación entre los clústeres, aunque con espacio para mejorar hacia el valor ideal de 1. Este resultado sugiere que, mientras que algunos clústeres están bien diferenciados, otros podrían tener solapamientos significativos o variaciones internas que reducen la claridad de la segmentación.

El **Índice Davies-Bouldin** proporcionó un valor de 1.4541, señalando una definición razonable de los clústeres. A pesar de que no se alcanza el valor mínimo posible, el índice sugiere una separación satisfactoria en comparación con la distancia intra-clúster.

Finalmente, el **Índice Calinski-Harabasz** mostró un valor robusto de 7311.53, lo que indica que los clústeres son relativamente densos y bien separados. Esta métrica, al ser especialmente sensible a la cohesión y separación de los clústeres, refuerza la conclusión de que la estructura de clústeres identificada es consistentemente significativa en el contexto de los datos analizados.

En conjunto, estas métricas respaldan la validez de los clústeres encontrados, aunque también sugieren que hay áreas para explorar mejoras en la segmentación. Sería beneficioso investigar la configuración del modelo y las características del conjunto de datos que podrían estar influyendo en la calidad de la agrupación observada.

Tabla 4.2. Metricas Clusterización GMM

Métrica	Valor
Silhouette Score	0,3117
Índice Davies-Bouldin	1,4541
Índice Calinski-Harabasz	7311,53

DBSCAN

Elección del número óptimo de clusters

Parámetro epsilon

La Figura 4.10 presenta un gráfico de la distancia K utilizado para determinar un valor apropiado para el parámetro *eps* del algoritmo DBSCAN. Los datos se ordenan por la distancia al quinto vecino más cercano, proporcionando una representación visual de la densidad del conjunto de datos. Se observa que la distancia se mantiene baja y bastante

uniforme para la mayoría de los puntos, lo que indica una alta densidad y proximidad entre las observaciones. Un aumento abrupto en la distancia al quinto vecino más cercano se identifica como el punto de codo (marcado con una línea vertical azul), sugiriendo un punto de transición natural entre los datos densamente y menos densamente poblados. Este punto de codo se considera una indicación del valor adecuado para *eps*, que define la escala de distancia máxima para que los puntos se consideren vecinos en el contexto de DBSCAN.

Una línea horizontal roja indica el percentil 95 de las distancias de la k-ésima vecindad más cercana, proporcionando un límite conservador para seleccionar *eps* y asegurar que el algoritmo se centre en el núcleo denso de los datos, minimizando así la inclusión de puntos de ruido en la formación de clústeres. La elección de este umbral se alinea con una estrategia de mitigación de anomalías, esencial para la precisión del análisis de clustering posterior. La metodología para determinar *eps* de esta manera subraya un enfoque empírico basado en la estructura inherente de los datos, crucial para la robustez de los clústeres identificados por DBSCAN.

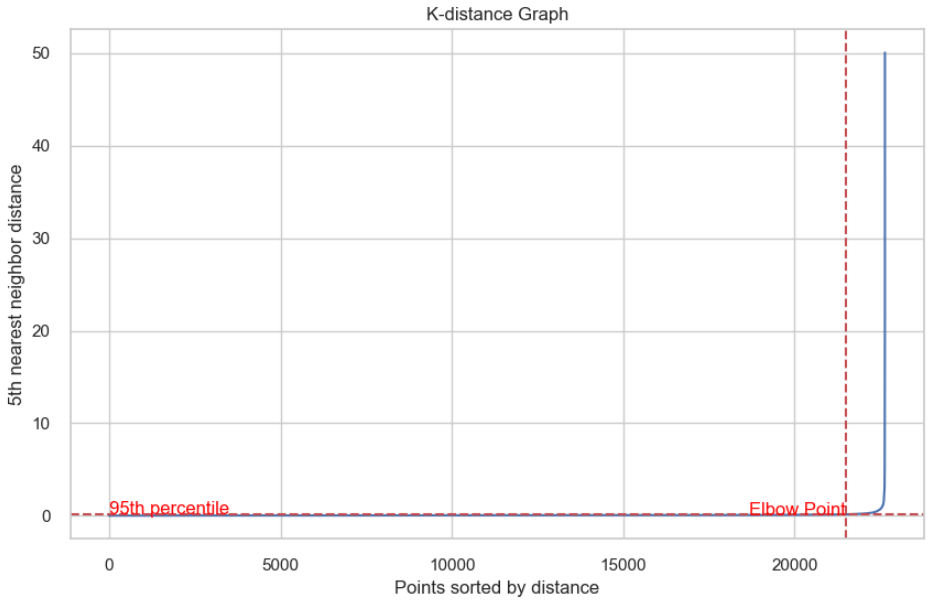


Figura 4.10. Identificación del Punto de Codo para la Estimación de Epsilon en DBSCAN

La Figura 4.11 muestra los resultados del clustering después de aplicar el algoritmo DBSCAN a nuestro conjunto de datos, que previamente se había transformado

utilizando PCA para reducir la dimensionalidad. Los dos primeros componentes principales, que representan las direcciones de máxima varianza en los datos, se utilizan como ejes para esta visualización bidimensional. Cada punto representa una observación, y el color asignado a cada uno indica su membresía a uno de los varios clústeres identificados por DBSCAN o su clasificación como ruido.

Se puede observar que DBSCAN ha identificado un total de siete clústeres (denotados por los colores del 0 al 6) que varían en tamaño y densidad. Además, una cantidad de puntos dispersos no se asignan a ningún clúster y se marcan como ruido, indicado por el color que representa 'Noise'. Estos son puntos que no cumplen con los criterios mínimos de densidad establecidos por los parámetros *eps* y *min_samples* de DBSCAN y, por lo tanto, se consideran anomalías en el contexto del conjunto de datos.

La capacidad de DBSCAN para distinguir entre clústeres de alta densidad y puntos de ruido es particularmente útil en aplicaciones donde la identificación de outliers es tan crítica como la detección de clústeres. Los clústeres identificados muestran una variedad de formas y tamaños, destacando la flexibilidad de DBSCAN en comparación con algoritmos de clustering que asumen clústeres de formas esféricas o de tamaños uniformes.

La presencia de clústeres claramente definidos y la identificación de outliers en la visualización subrayan la complejidad del conjunto de datos analizado y la eficacia del enfoque de clustering basado en densidad para discernir estructuras subyacentes no lineales y no convencionales en los datos.

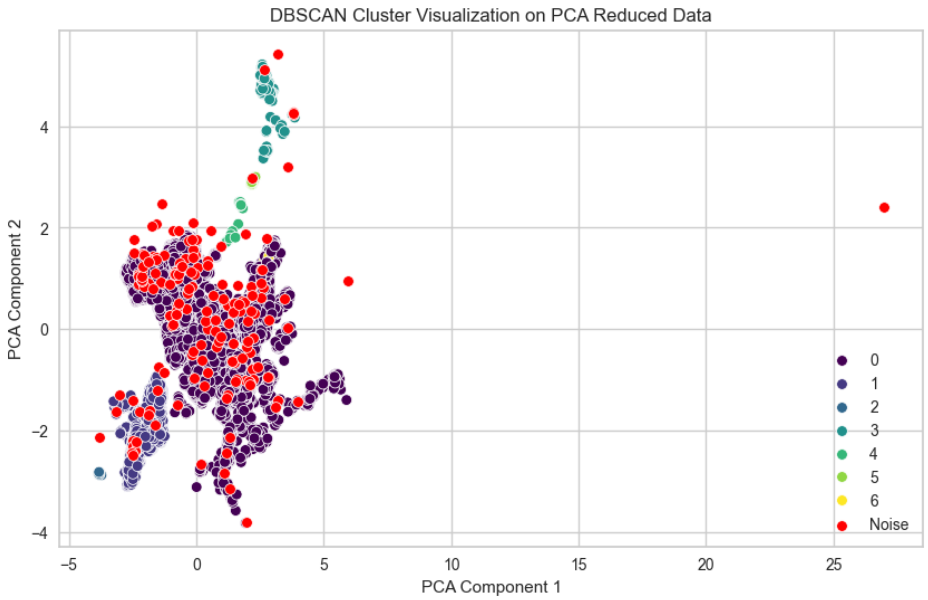


Figura 4.11. Distribución de Clusters con DBSCAN en Componentes Principales de Análisis PCA

Tras la aplicación del algoritmo DBSCAN al conjunto de datos dimensionalmente reducido a través de PCA, se calcularon varias métricas de validación de clustering para cuantificar la cohesión y separación de los clústeres identificados.En la tabla 4.3 se puede observar el **Silhouette Score** obtenido fue de 0.1845, lo cual es relativamente bajo y sugiere que hay un margen de mejora en términos de la cohesión interna de los clústeres y su separación con respecto a otros clústeres. Este valor indica que, aunque algunos clústeres están bien definidos, otros pueden estar superpuestos o contener puntos que podrían estar mejor situados en clústeres adyacentes.

El **Índice Davies-Bouldin**, con un valor de 0.6109, indica una separación razonable entre los clústeres. Este índice favorece valores más bajos, con un valor de cero siendo el ideal, lo que implica que los clústeres están bien separados y definidos. Por lo tanto, a pesar de la baja puntuación de Silhouette, el valor del Índice Davies-Bouldin sugiere que los clústeres no están excesivamente entrelazados y que cada clúster es relativamente distinto de los demás.

Finalmente, el **Índice Calinski-Harabasz** proporcionó un valor de 1475.8, que es una medida de la dispersión entre los clústeres comparada con la dispersión dentro de los clústeres. Un valor más alto indica clústeres bien definidos y separados. Este valor

relativamente alto sugiere que los clústeres son bastante compactos y bien separados en comparación con la variación dentro de los clústeres, lo cual es favorable para la interpretación de los resultados del clustering.

En conjunto, estas métricas proporcionan una visión mixta de la calidad del clustering. El bajo valor de Silhouette podría ser un reflejo de las características únicas del algoritmo DBSCAN, que no fuerza a todos los puntos a ser asignados a un clúster, permitiendo en su lugar que algunos puntos sean clasificados como ruido. Esto puede ser particularmente útil en conjuntos de datos con outliers o con estructuras de clústeres no convencionales. Por otro lado, los valores de los Índices Davies-Bouldin y Calinski-Harabasz sugieren que los clústeres identificados tienen una buena separación y definición. La discrepancia entre el Silhouette Score y los otros índices podría deberse a la presencia de ruido y a la variabilidad en la densidad de los clústeres, características que son manejadas de manera inherente por DBSCAN.

Estos resultados resaltan la importancia de seleccionar métricas de evaluación que sean coherentes con los objetivos del análisis de clustering y las características del algoritmo utilizado. Además, subrayan la relevancia de considerar múltiples perspectivas de evaluación para obtener una interpretación holística de la calidad del clustering.

Tabla 4.3. Metricas Clusterización DBSCAN

Métrica	Valor
Silhouette Score	0,1845
Índice Davies-Bouldin	0,6109
Índice Calinski-Harabasz	1475,80

4.3 Creacion de canastas de productos - Fp Growth

La Figura 4.12 ilustra los diez conjuntos de ítems más frecuentes en el conjunto de datos, evaluados por su soporte en el contexto de un análisis de mercado de cestas de compras. Los resultados revelan que el ítem con el identificador ’5640’ es el más prevalente entre todas las transacciones, lo que indica su alta relevancia para los consumidores en el conjunto de datos. Los conjuntos de ítems que involucran los identificadores ’320’ y ’2830’

también muestran un soporte significativo, destacando su importancia en el conjunto de datos analizado.

Interesantemente, los conjuntos de múltiples ítems como {'40', '30', '20'} presentan un soporte considerable, lo que sugiere que estos ítems tienden a ser comprados juntos con una frecuencia notable. Este tipo de información es crucial para el diseño de estrategias de marketing, como la colocación de productos y las promociones cruzadas, ya que indica patrones de compra conjunta.

La variación en el soporte entre los distintos conjuntos de ítems también subraya la diversidad en el comportamiento de compra de los consumidores. Tales insights permiten a los minoristas y a los analistas de datos comprender mejor las tendencias de compra y optimizar las recomendaciones de productos y el inventario en consecuencia.

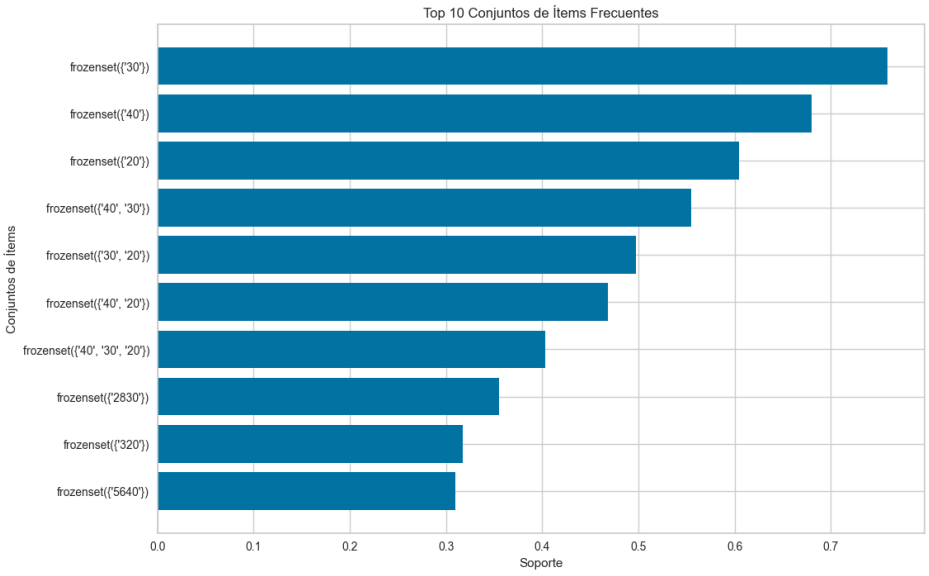


Figura 4.12. Soporte de Conjuntos de Items Frecuentes

La figura 4.13 ilustra la relación entre la confianza y el lift para las reglas de asociación derivadas del conjunto de datos. Los contornos de densidad resaltan las regiones donde se concentran las reglas, con una notable acumulación de reglas que presentan una confianza de entre 0.6 y 0.9 y un lift de entre 2 y 6. Estos resultados sugieren que hay un número significativo de reglas fuertes en el conjunto de datos que indican asociaciones positivas entre los ítems involucrados. La línea de base punteada marcada como 'Lift Base' atraviesa el eje Y en lift = 1, subrayando el punto de referencia

de independencia. Los valores de lift superiores a esta línea base demuestran la tendencia de ciertos ítems a co-ocurrir en transacciones más frecuentemente de lo que se esperaría por casualidad, lo cual es indicativo de patrones de compra significativos y potencialmente accionables para estrategias de marketing y colocación de productos. La distribución de los puntos sugiere que mientras un gran número de reglas exhibe una confianza y un lift moderados a altos, hay una variedad de reglas que exhiben una confianza extremadamente alta pero con un rango de lift más amplio, lo que merece una investigación adicional para entender las particularidades de estas asociaciones.

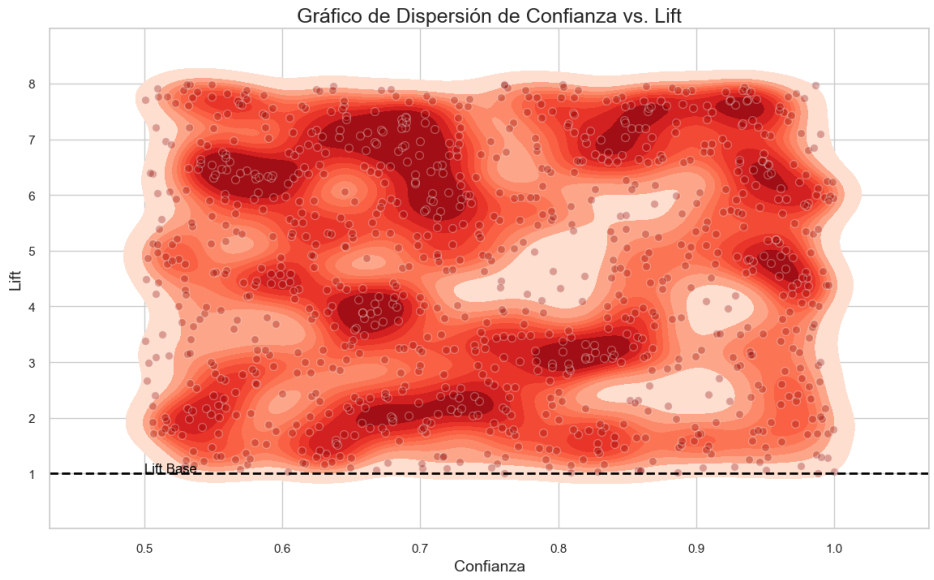


Figura 4.13. Mapa de Calor de la Relación entre Confianza y Lift en Reglas de Asociación

La Figura 4.14 ilustra la red de reglas de asociación perteneciente al Clúster 0, compuesta por nodos que representan conjuntos de ítems y aristas que indican las reglas de asociación entre estos. Los nodos están marcados con identificadores únicos, como 2040, 310, 1650, etc., que podrían corresponder a ítems individuales o a transacciones específicas dentro del conjunto de datos. Las aristas entre los nodos están anotadas con dos métricas clave: la confianza (C) y el *lift* (L). La confianza es una medida de la probabilidad condicional de que un ítem B sea comprado dado que se ha comprado un ítem A, y el lift es una medida de la fuerza de una regla sobre la base de la frecuencia con la que ítems A y B ocurren juntos, más de lo esperado por su respectiva frecuencia individual.

La arista etiquetada como ‘C:0.75, L:1.40’ sugiere que la regla de asociación correspondiente es significativa, con una confianza del 75% y un lift de 1.40, indicando que la presencia de los ítems en el antecedente incrementa en un 40% la probabilidad de ocurrencia del ítem en el consecuente sobre la expectativa de independencia. Este nivel de detalle permite una interpretación rica y específica de las relaciones entre ítems, subrayando no solo la frecuencia de las asociaciones, sino también su relevancia y robustez.

La visualización en red facilita la comprensión de la complejidad y la estructura de las relaciones entre los ítems, destacando los nodos más conectados como posibles elementos centrales en el clúster. Estos pueden ser de particular interés para estrategias de marketing enfocadas en la colocación de productos o en la recomendación de ítems relacionados. Este tipo de análisis es vital para la extracción de conocimiento accionable a partir de grandes conjuntos de datos de transacciones.

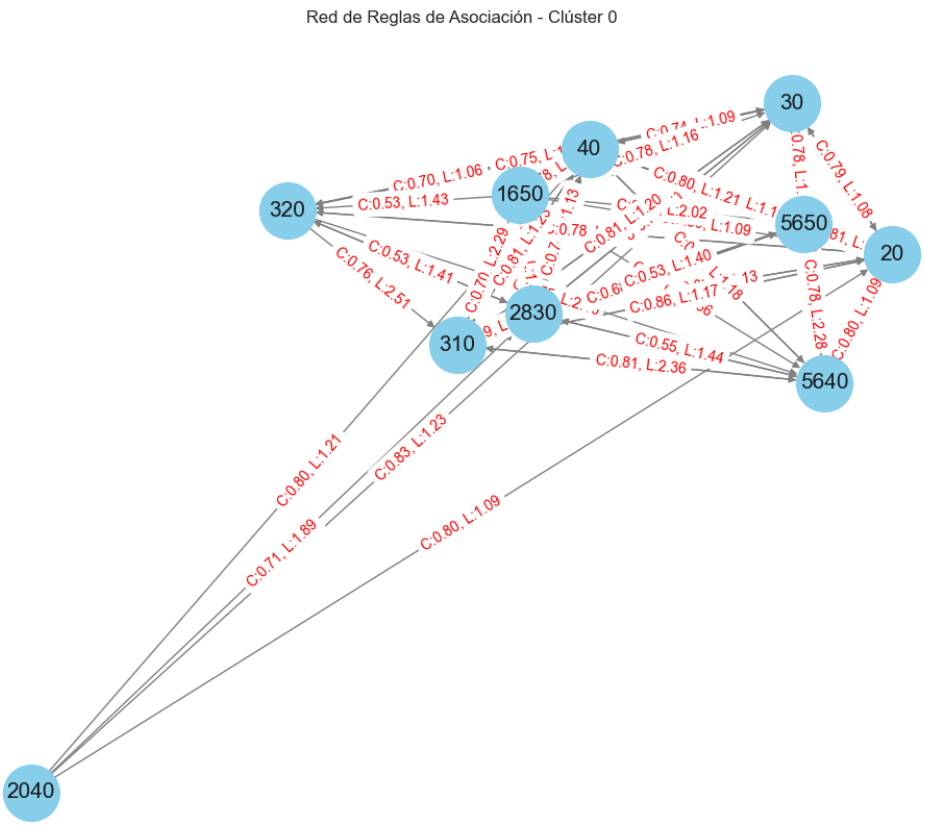


Figura 4.14. Grafo de Interconexión de Reglas de Asociación con Confianza y Lift Cluster 1

La figura 4.15 representa una red de asociación para el Clúster 1, compuesta por múltiples elementos interrelacionados. Los nodos, diferenciados por tamaño, denotan la importancia relativa de cada elemento, mientras que las conexiones entre ellos, anotadas con valores de confianza (C) y significancia (L), cuantifican la fortaleza de las asociaciones. La red exhibe una estructura compleja con varios nodos altamente conectados, lo que sugiere una interdependencia significativa entre los elementos del clúster. Esta topología puede indicar áreas de influencia central dentro del clúster o puede usarse para identificar subgrupos de elementos estrechamente vinculados. La interpretación detallada de los valores de confianza y significancia de las conexiones podría proporcionar *insights* adicionales sobre la cohesión del clúster y la relevancia de las asociaciones específicas.

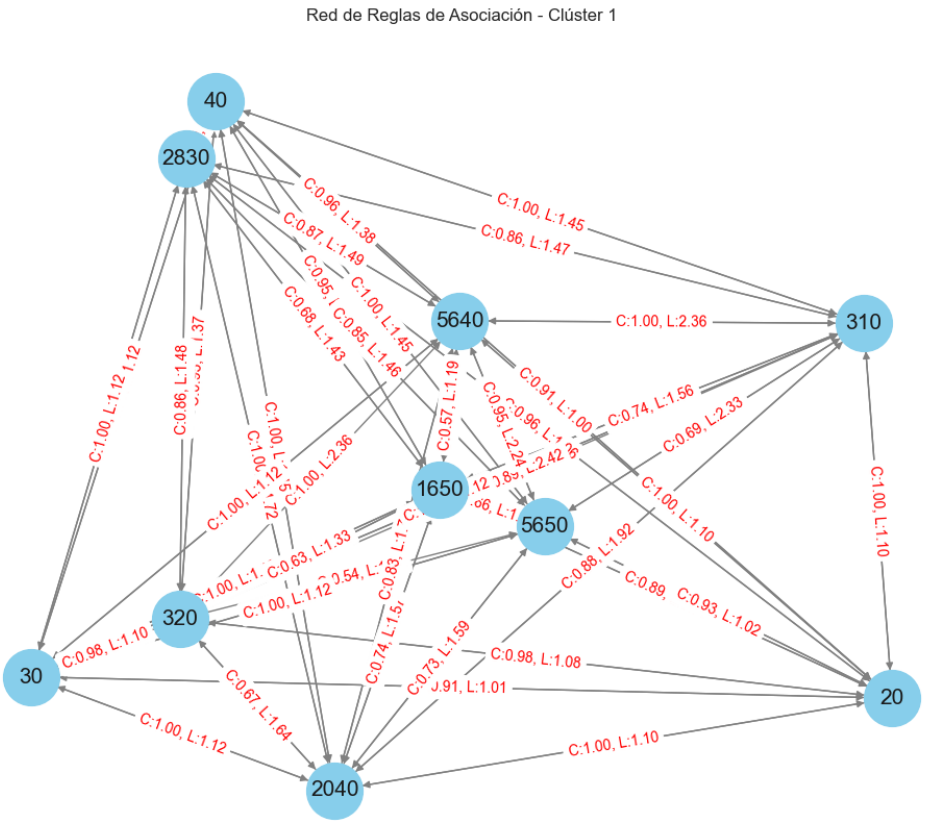


Figura 4.15. Grafo de Interconexión de Reglas de Asociación con Confianza y Lift Cluster 2

La figura 4.16 ilustra la estructura de la red de asociación para el Clúster 2. Los nodos, cuyo tamaño varía, representan elementos individuales dentro del clúster, y sus

tamaños relativos sugieren la centralidad o importancia de cada uno dentro de la red. Las conexiones entre nodos están etiquetadas con métricas de *confianza* y *soporte* (o *significancia*), las cuales indican la robustez y relevancia de las relaciones respectivas. Nodos clave, como el 5640 y el 2830, actúan como centros neurálgicos con numerosas conexiones, posiblemente denotando elementos con un alto grado de influencia o interacción dentro del clúster. La distribución de los valores de confianza y significancia a lo largo de la red proporciona una visión cuantitativa sobre la estructura y dinámica de las relaciones inter-elementos, lo que es esencial para entender la cohesión interna del clúster y para identificar subestructuras significativas.

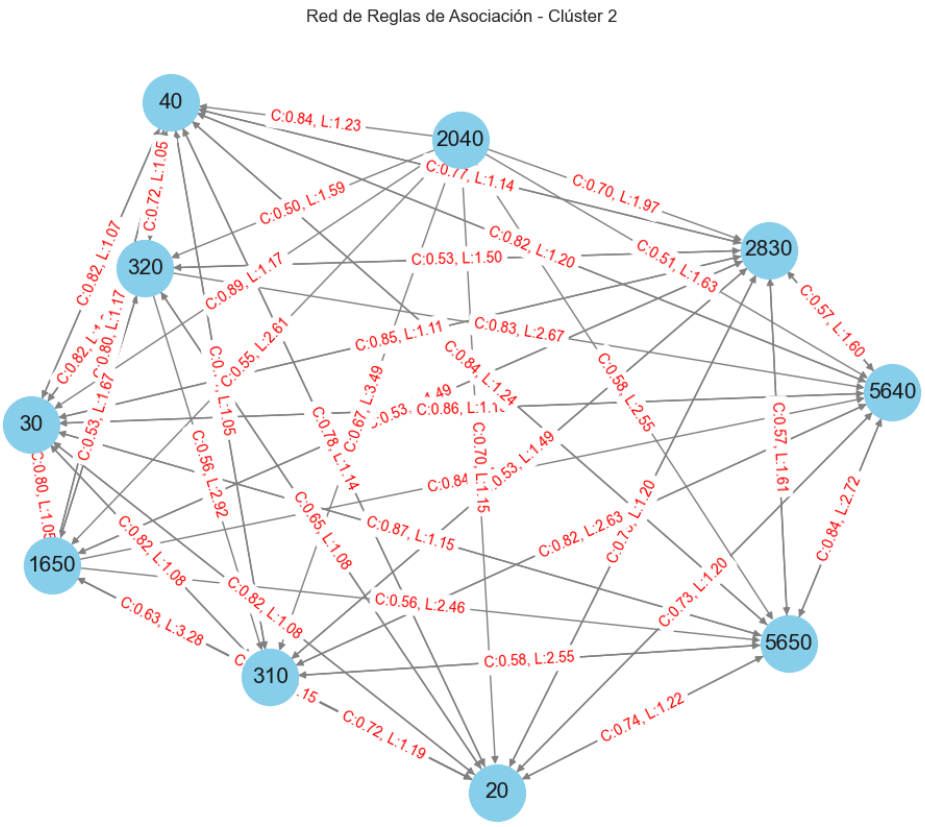


Figura 4.16. Grafo de Interconexión de Reglas de Asociación con Confianza y Lift Cluster 3

La figura 4.17 muestra la red de asociación para el Clúster 3, donde los nodos representan elementos y sus conexiones indican la presencia de asociaciones significativas entre ellos. Estos nodos varían en tamaño, reflejando su centralidad y la cantidad de conexiones en la red. Las líneas que interconectan los nodos llevan

etiquetas con valores de *confianza* y un segundo parámetro, posiblemente *soporte* o *levantamiento*, proporcionando una medida cuantitativa de la fortaleza de cada asociación. Observamos que ciertos nodos, como el 2830 y el 5640, se destacan como *hubs* centrales dentro de la red, lo que sugiere que son elementos clave dentro del clúster con numerosas conexiones. Estas características de la red podrían ser cruciales para identificar patrones de asociación, entender la estructura y dinámica del clúster, y potencialmente para descubrir subgrupos con comportamientos o propiedades similares.

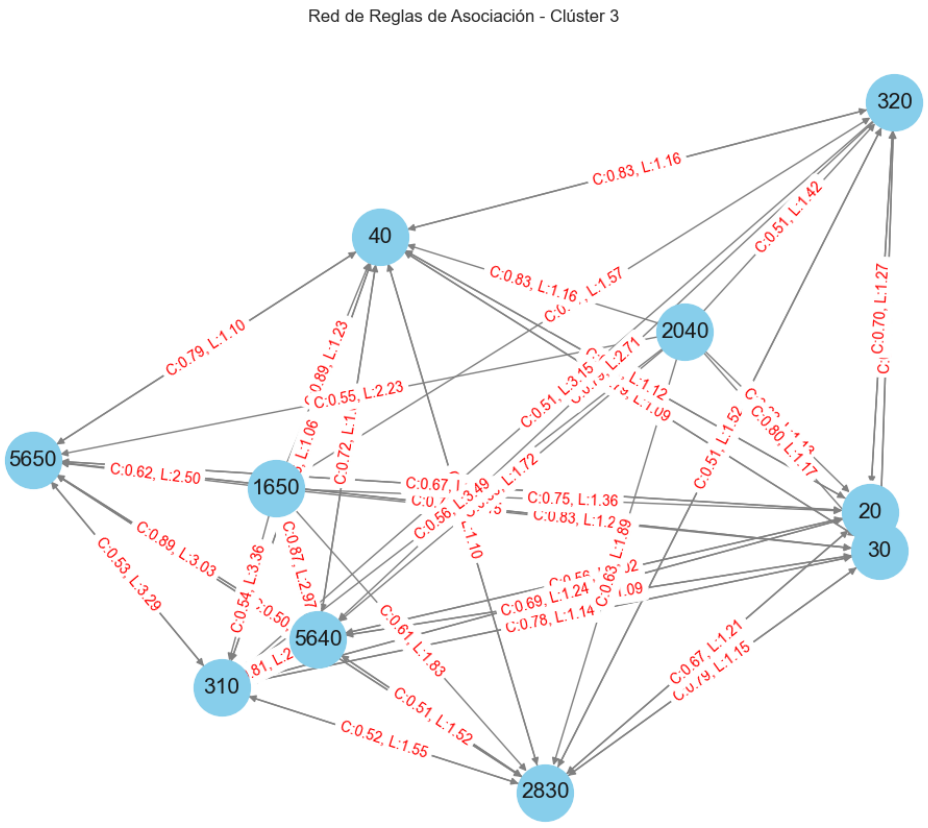


Figura 4.17. Grafo de Interconexión de Reglas de Asociación con Confianza y Lift Cluster 4

La figura 4.18 representa la red de asociación para el Clúster 4, ilustrando la interconexión entre los elementos del clúster. Los nodos de diferentes tamaños indican variaciones en la centralidad y el número de conexiones. Las líneas que unen los nodos llevan anotaciones de *confianza* y *significancia*, revelando la solidez de las asociaciones. Nodos con muchas conexiones, como el 1650 y el 320, destacan como

elementos significativos dentro de la red, posiblemente actuando como concentradores de la actividad del clúster. La variabilidad en los valores de confianza y significancia sugiere una diversidad en la fuerza de las relaciones, lo que puede ser clave para comprender la estructura y dinámica del clúster. El análisis de estas conexiones proporciona una perspectiva cuantitativa sobre la cohesión del clúster y permite la identificación de subgrupos significativos dentro de la red.

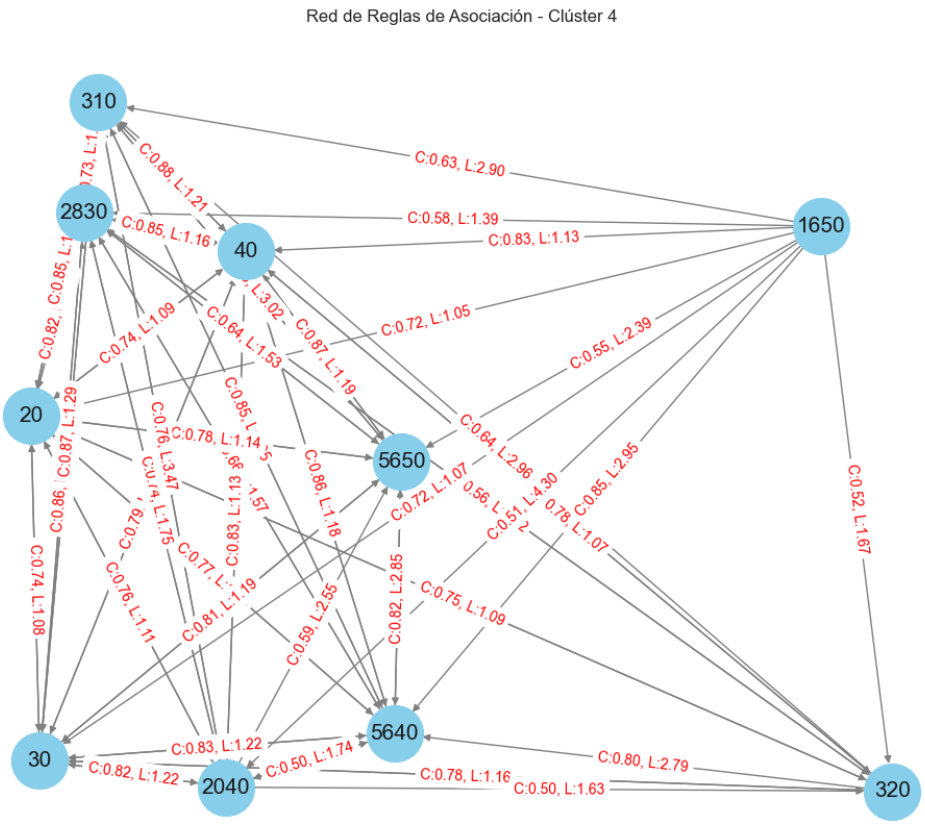


Figura 4.18. Grafo de Interconexión de Reglas de Asociación con Confianza y Lift Cluster 5

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

- El análisis del codo determinó que cinco clústeres minimizan la distorsión, con un cambio sustancial en la pendiente de la curva de distorsión alrededor de $k = 5$. Este punto refleja un equilibrio entre la minimización de la varianza intra-clúster y la maximización de la diferenciación entre clústeres, proporcionando una justificación empírica para la selección del número de clústeres en el contexto del conjunto de datos estudiado.

- La ejecución de K-Means en el espacio reducido por PCA identificó cinco clústeres con una distribución heterogénea en términos de densidad y ubicación espacial. Esta diversidad es indicativa de la variabilidad en las características subyacentes del conjunto de datos, lo que demuestra la capacidad de K-Means para capturar y resumir la estructura compleja dentro de los datos.

- Evaluación Integral de Clústeres: La aplicación de métricas consolidadas de evaluación de clústeres ha proporcionado una vista comprensiva de la calidad del agrupamiento. Un Silhouette Score de 0.4693, aunque distante del ideal, sugiere una razonable cohesión intra-clúster y una separación inter-clúster, mientras que el índice Calinski-Harabasz de 14276.37 refuerza la distinción entre clústeres al indicar una dispersión inter-clúster significativa comparada con la dispersión intra-clúster.

  K-Means:

  - Silhouette Score: 0.46 (relativamente alto, indica buena separación y cohesión).
  - Davies-Bouldin Score: 0.74 (medio, indica una separación aceptable).

- Calinski-Harabasz Score: 14276.37 (muy alto, indica clusters muy densos y bien separados).

Gaussian Mixture Model (GMM):

- Silhouette Score: 0.31 (más bajo que K-Means, indica una peor separación y cohesión).
- Davies-Bouldin Score: 1.45 (más alto que K-Means, indica una peor separación entre los clusters).
- Calinski-Harabasz Score: 7311.53 (menor que K-Means, indica menos densidad y separación).

DBSCAN:

- Silhouette Score: 0.18 (el más bajo, lo que puede indicar una separación y cohesión inadecuadas).
- Davies-Bouldin Score: 0.61 (el más bajo, indica la mejor separación relativa entre los clusters).
- Calinski-Harabasz Score: 1475.80 (el más bajo, indica la menor densidad y separación).

En base a estas métricas, **K-Means** parece ser el mejor método para tu conjunto de datos, ya que tiene el *Silhouette Score* más alto, un *Davies-Bouldin Score* medio y el *Calinski-Harabasz Score* más alto, lo que indica que los clusters formados son densos, bien separados y tienen una buena cohesión. Sin embargo, la elección del método de clustering también puede depender de la naturaleza de tus datos y del contexto específico de tu análisis.

- El enfoque de FP Growth ha resaltado conjuntos de ítems con soportes notables, particularmente el ítem ’5640’, que sobresale en frecuencia de compra. Este patrón sugiere su centralidad en las preferencias de los consumidores y su potencial como punto de enfoque para estrategias de marketing y promoción cruzada.

- Las reglas de asociación desenterradas con elevados niveles de confianza y lift revelan conexiones fuertes y estadísticamente significativas entre pares de ítems. Estas asociaciones ofrecen una ventana hacia comportamientos de compra coherentes que pueden ser explotados para promociones dirigidas y para aumentar la efectividad de las recomendaciones de productos.
- La red de reglas de asociación, ilustrada visualmente, demuestra la complejidad y la riqueza de las relaciones entre ítems. Los ítems con múltiples conexiones de alta confianza y lift actúan como nodos centrales en la red, lo que refleja su importancia estratégica y su potencial para influir en decisiones de compra cruzadas.
- Los insights derivados de la red de asociaciones y los patrones de compra deben ser utilizados para informar decisiones de marketing. Estos hallazgos pueden guiar la colocación de productos y el diseño de promociones para maximizar la venta cruzada y la satisfacción del cliente.
- Se propone una exploración más profunda de la configuración del modelo de clústering y las características detalladas de las reglas de asociación. Esto puede incluir la optimización de hiperparámetros y la validación de las reglas en distintos contextos de mercado para perfeccionar aún más la segmentación y la precisión de las predicciones de compra.
- El estudio aporta una metodología cuantitativa rigurosa para la evaluación de clústeres y asociaciones en grandes bases de datos. Los métodos utilizados expanden el conocimiento práctico y teórico en el campo de la minería de datos, especialmente en la aplicación de clústering no supervisado y reglas de asociación en entornos de big data.
- Los resultados subrayan la importancia de la ciencia de datos en la formulación de tácticas de negocio basadas en evidencia. La interpretación de patrones complejos de compra es crucial para el desarrollo de estrategias de marketing más efectivas y para el aumento de la rentabilidad.
- Aunque el Silhouette Score indica clústeres bien formados, también sugiere la

presencia de solapamientos, señalando la necesidad de una segmentación más precisa y quizás un enfoque de clústering más sofisticado para diferenciar más claramente entre subgrupos similares.

- El tiempo de ajuste del modelo aumenta levemente con el número de clústeres, pero se mantiene manejable, lo cual es un punto favorable desde la perspectiva de la eficiencia computacional, permitiendo la escalabilidad del modelo a conjuntos de datos más grandes sin una penalización significativa en el rendimiento.

- El análisis BIC para el modelo de mezcla gaussiana muestra una disminución pronunciada en los valores al incrementar el número de clústeres, con un punto de inflexión en 10 clústeres. Esto indica un modelo GMM con un ajuste óptimo que equilibra la complejidad del modelo y la representación precisa de la estructura de datos.

- La determinación del parámetro epsilon para DBSCAN a través del gráfico K-distance proporciona un criterio empírico para elegir un valor de eps que refleje la densidad inherente del conjunto de datos y minimice la inclusión de puntos de ruido.

- El algoritmo DBSCAN demuestra su capacidad para identificar clústeres de alta densidad y aislar puntos de ruido, lo que es particularmente valioso en aplicaciones donde es crucial distinguir entre datos típicos y atípicos.

- La combinación de métricas utilizadas para evaluar el clustering de DBSCAN ofrece una visión integral de la calidad del agrupamiento, destacando tanto las fortalezas como las áreas para mejorar la precisión del modelo.

- Las métricas como el Silhouette Score de 0.1845, el Índice Davies-Bouldin de 0.6109 y el Índice Calinski-Harabasz de 1475.80 proporcionan una imagen matizada de la efectividad del clustering, señalando que hay una estructura de clúster significativa pero también destacando la presencia de ruido y la variabilidad en la densidad de los clústeres.

- El análisis de los diez conjuntos de ítems más frecuentes ofrece una comprensión clara de los patrones de consumo, resaltando ítems específicos y combinaciones de ítems que son significativamente más comunes en las transacciones de los consumidores.
- Los métodos de clustering y análisis de reglas de asociación han demostrado ser efectivos para descubrir estructuras complejas y patrones significativos en el comportamiento de compra. Las métricas consolidadas respaldan la robustez del esquema de clustering implementado y destacan oportunidades para el perfeccionamiento del modelo.

BIBLIOGRAFÍA

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. pages 487–499.

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms.*

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4:126.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning. Springer.* Springer.

Borgelt, C. (2005). An implementation of the fp-growth algorithm. pages 1–5.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, pages 1–27.

Das, P. and Chaudhury, S. (2007). Prediction of retail sales of footwear using feedforward and recurrent neural networks. *Neural Computing and Applications*, 16:491–502.

Davies, D. L. and Bouldin, D. W. (1974). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 224–227.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *ournal of the Royal Statistical Society: Series B (Methodological)*, 39:1–38.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Hahsler, M., Grün, B., and Hornik, K. (2005). Arules – a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, pages 14–15.

Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. 3ra ed. edition.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *International Conference on Management of Data,*, pages 1–12.

Jang, J. and Jiang, H. (2018). Dbscan++: Towards fast and scalable density clustering.

Larose, D. T. and Larose, C. D. (2014). Discovering knowledge in data an introduction to data mining second edition wiley series on methods and applications in data mining.

Lu, C. J., Shao, Y. E., and Li, C. C. (2014). Recognition of concurrent control chart patterns by integrating ica and svm. *Applied Mathematics and Information Sciences*, 8:681–689.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.

Martínez, L. and Fernández, A. (2021). Cluster analysis in consumer segmentation: A marketing strategy. *Journal of Marketing Strategies*, 33:112–129.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley Sons.

Olson, N. (2010). *Taken for granted : the construction of order in the process of library management system decision making*. Valfrid.

Rasmussen, C. E. (2000). The infinite gaussian mixture model. *Advances in Neural Information Processing Systems*, pages 554–560.

Reynolds, D. A. (2015). Gaussian mixture models. *Encyclopedia of Biometrics*, pages 827–832.

Rodríguez-Cruz, Y. and Pinto, M. (2018). Modelo de uso de información para la toma de decisiones estratégicas en organizaciones de información. *Transinformacao*, 30(1):51–64.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Savasere, A., Omiecinski, E., and Navathe, S. N. (1995). An efficient algorithm for mining association rules in large databases. pages 432–444.

Seetharaman, A., Niranjan, I., Tandon, V., and Saravanan, A. S. (2016). Impact of big data on the retail industry.

Ultsch, A. and Lötsch, J. (2022). Euclidean distance-optimized data transformation for cluster analysis in biomedical data (edotrans). *BMC Bioinformatics*, 23(1):Article 233.

Walker, K. B. and Mcclellandt, L. A. (1991). Management forecasts and statistical prediction model forecasts in corporate budgeting. *Journal of Accounting Research*, 29.

Wedel, M. and Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations. *Kluwer Academic Publishers*.

Zhou, Y. and Wang, X. (2022). Leveraging machine learning in customer cluster models: Uncovering hidden segments. 39:456–473.